PB85-236891

Summary of Complex Terrain Model Evaluation

American Meteorological Society, Boston, MA

Prepared for

Environmental Protection Agency
Research Triangle Park, NC

Jul 85

U.S. Department of Commerce
National Technical Information Service
NTIS

PB85-236891

EPA/600/3-85/060
July 1985

SUMMARY OF COMPLEX TERRAIN MODEL EVALUATION

by

Fred D. White, Editor
American Meteorological Society
45 Beacon Street
Boston, Massachusetts 02108

and

Jason K. S. Ching, Robin L. Dennis, and William H. Snyder
Meteorology and Assessment Division
Atmospheric Sciences Research Laboratory
Research Triangle Park, North Carolina 27711

Project Officer

Francis A. Schiermeier
Meteorology and Assessment Division
Atmospheric Sciences Research Laboratory
Research Triangle Park, North Carolina 27711

ATMOSPHERIC SCIENCES RESEARCH LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27711

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA/600/3-85/060	2.	3. RECIPIENT'S ACCESSION NO. PB85 236891 /AS
4. TITLE AND SUBTITLE SUMMARY OF COMPLEX TERRAIN MODEL EVALUATION		5. REPORT DATE July 1985
		6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Fred D. White, Jason K..S. Ching, Robin L. Dennis, and William H. Snyder		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS American Meteorological Society, Boston, MA Meteorology and Assessment Division Research Triangle Park, NC 27711		10. PROGRAM ELEMENT NO. CDWA1A/02 0279 (FY-85)
		11. CONTRACT/GRANT NO. CR 810297 and Inhouse
12. SPONSORING AGENCY NAME AND ADDRESS Atmospheric Sciences Research Laboratory -- RTP, NC Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, NC 27711		13. TYPE OF REPORT AND PERIOD COVERED Interim (FY-84/85)
		14. SPONSORING AGENCY CODE EPA/600/09

15. SUPPLEMENTARY NOTES

16. ABSTRACT

The Environmental Protection Agency conducted a scientific review of a set of eight complex terrain dispersion models. TRC Environmental Consultants, Inc. calculated and tabulated a uniform set of performance statistics for the models using the Cinder Cone Butte and Westvaco Luke Mill data bases. Three members of the EPA Meteorology and Assessment Division reviewed the performance statistics and presented objective analyses of the models and their performance. An American Meteorological Society Steering Committee summarized the reviews and formulated three conclusions: (1) none of the models can be regarded as up-to-date scientifically; (2) one model exhibited much better performance statistics than did the others; and (3) overprediction was the most common problem with the models. This report consists of the AMS summary and copies of three independent reviews conducted to evaluate the model performance.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS *(This Report)* UNCLASSIFIED	21. NO. OF PAGES 131
	20. SECURITY CLASS *(This page)* UNCLASSIFIED	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

NOTICE

The information in this document has been funded by the United States Environmental Protection Agency under Cooperative Agreement 810297 to the American Meteorological Society. It has been subject to the Agency's peer and administrative review, and it has been approved for publication as an EPA document.

Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

ABSTRACT

The Environmental Protection Agency conducted a scientific review of a set of eight complex terrain dispersion models. TRC Environmental Consultants, Inc. calculated and tabulated a uniform set of performance statistics for the models using the Cinder Cone Butte and Westvaco Luke Mill data bases. Three members of the EPA Meteorology and Assessment Division reviewed the performance statistics and presented objective analyses of the models and their performance. An American Meteorological Society Steering Committee summarized the reviews and formulated three conclusions: (1) none of the models can be regarded as up-to-date scientifically; (2) one model exhibited much better performance statistics than did the others; and (3) overprediction was the most common problem with the models. This report consists of the AMS summary and copies of three independent reviews conducted to evaluate the model performance.

CONTENTS

Abstract. iii

1. Summary of Complex Terrain Model Evaluation 1
   AMS Steering Committee

2. Review of Complex Terrain Model Performance 9
   Jason K. S. Ching

3. Review of Complex Terrain Model Performance 61
   Robin L. Dennis

4. Review of Complex Terrain Model Performance 113
   William H. Snyder

SUMMARY OF COMPLEX TERRAIN MODEL EVALUATION

I. SUMMARY

The complex terrain model evaluation is the third in the series conducted under the cooperative agreement between the American Meteorological Society (AMS) and the U.S. Environmental Protection Agency (EPA). Like the earlier reviews of rural and urban models, this evaluation included three distinct tasks, a performance evaluation in which model predictions and field data were compared, a set of independent peer reviews, and a brief summary of the evaluation developed by the AMS members of the Committee.

This evaluation differed from the previous projects in two respects. First, the peer reviewers consisted of three EPA staff members, working independently. Second, the performance evaluation included two data sets instead of one.

All of the reviewers were on assignment from NOAA to the Meteorology and Assessment Division, Atmospheric Sciences Research Laboratory:

Jason K.S. Ching
Robin L. Dennis
William H. Snyder

In our judgment, the reviewers presented valuable, objective studies of the models and their performance.

They mirrored the reactions of the reviewers of the rural and urban models and the AMS Committee members in being disappointed in the technical quality of the models, and in concluding that a massive statistical analysis is not the best way to analyze model performance. The difficulty with the latter is that one obtains much numerical information about model performance on limited data sets, but very little about the strengths and weaknesses of the models as they deal with specific meteorological problems. This group of reviewers and the Committee members believe that future evaluations should stress understanding of the physics of the models and study of their performance in specific case studies, with less emphasis on massive statistical efforts.

It is important to keep in mind that the current evaluation developed data on rather limited aspects of transport and diffusion in complex terrain. The models were originally intended to deal with flow and dispersion around isolated obstacles, in particular on the windward side of such obstacles. Thus, the performance data addressed only the source-receptor relationships on the upslope and crest portions of nearby terrain obstacles. This information is of course important in regulatory applications, but it represents a very limited aspect of complex terrain problems. It tells one nothing about up-and down-valley flow, slope flow or effects beyond the immediate terrain obstacles.

However, within these limitations, three conclusions about the models and their performance are evident:

- **The Models are Scientifically Outdated**

None of the models can be described as up-to-date scientifically. Seven of the eight models still utilize the Gaussian diffusion kernel with simple modifications to account for the terrain effects. The key modification is the reduction of the effective stack height of the plume by some fraction of the terrain height to simulate the passage of the plume closer to the terrain. The eighth model develops a wind field and treats diffusion differently, but it is difficult to evaluate from the information presented in the users guide.

- **RTDM a Clear Choice**

This performance evaluation was unlike either the rural or the urban reviews in that one model, RTDM, exhibited much better performance statistics than the others. In both data sets, this model predicted concentrations that were quite close to the observed, on average, although there were indeed large deviations between individual observations and predictions. In contrast, the other models exhibited large deviations even in the average ratio of predicted-to-observed concentrations.

RTDM is also one of the more up-to-date models, including improvements that are missing in most of the others. Like COMPLEX/PFM, it determines the dividing streamline height, H crit., during stably stratified flow conditions, treating diffusion above and below this elevation differently. It also prevents the unrealistic increase of the crosswind integrated concentration with distance that would violate the second law of thermodynamics. The model accounts for wind shear and utilizes onsite meteorological data in detail.

- **Overprediction is the Rule**

One can generalize in saying that overprediction is the most common problem with the models, and frequently the overprediction is sufficiently gross as to make regulatory application meaningless. This overprediction seems to be associated largely with plume impingement on terrain being predicted to occur where it does not, or in calculating excessive concentrations when it does, but it is dangerous to generalize about the reasons for poor performance.

II. MODELS

Eight models were included in the evaluation:

COMPLEX I
COMPLEX II
4141
PLUME 5
RTDM
SHORTZ
COMPLEX/PFM
IMPACT

Five of the models listed are basically simple Gaussian models, using minor adaptations to deal with some aspects of complex terrain flow and dispersion. These adaptations include modified terrain-receptor height adjustments to account for vertical plume motion over obstacles, enhanced lateral dispersion under certain circumstances, adjustments to account for wind direction shear, and limitation of plume approach height on impaction to some minimum value. RTDM and COMDPLEX/PFM are more contemporary in a scientific sense, the latter using potential flow adjustments in the wind field for part of the calculations, but none of these seven develops or simulates the three-dimensional transport and diffusion field that is actually present in such situations. Of the group evaluation, only IMPACT treats the problem with such a technique, and it is difficult in a review such as this one to determine how faithful a depiction this may be.

Inasmuch as the technical details of the models have been well summarized by TRC, we did not consider it necessary to repeat the description here. Similarly, since the reviews themselves are contained in this document, there is no point in providing a digested version of them.

III. DATA SETS

Two data sets were used for the model evaluation, representing a departure from the rural and urban reviews that were completed earlier. This is of course commendable, since it does not leave the evaluation at the mercy of a single batch of data. However, although the two data sets are probably the best that could have been used for the validation study, they leave much to be desired in terms of completeness and representativeness.

The Cinder Cone Butte data were the product of an excellent research project. The key terrain feature is a nearly perfect cone rising from a nearly flat plain, and both the concentration data and meteorological measurements were as detailed as one could possibly hope to achieve. The sources, however, consisted of passive releases of tracer gases and smoke which, while they provided very reliable concentration data, did not represent typical industrial sources in which large buoyant plumes are common. Thus, this set of data is unrepresentative in a very important way.

The second data set, obtained at the Westvaco plant site between 1979 and 1981, represented a much more typical industrial problem, but the concentration and meteorological data obtained were limited. Furthermore, one could hardly describe this particular site as being ideal from the standpoint of a validation study. The plant is situated in a rather winding valley, a configuration which distorted both the flow trajectories and the diffusion from an idealized complex terrain problem.

It was necessary to condense both data sets somewhat to make them digestible and consistent for this evaluation. It was also necessary to limit the study of one of the models, IMPACT, to considerably less than the full year of Westvaco data, presumably because of computer costs.

IV. PERFORMANCE

As we have noted earlier, the models tended to overpredict very seriously in most applications. One could choose any one of a number of statistical summaries to illustrate the point, but the following provides the flavor of the results:

Average Ratios of the 25 Highest One-Hour Predictions to the 25 Highest Observations Without Regard to Time or Location

	Ratios	
Model	WESTVACO SO_2	CINDER CONE TRACER
COMPLEX I	9.2	1.6
COMPLEX II	19.6	3.8
4141	6.2	3.0
RTDM	1.7	1.0
PLUME 5	7.4	2.7
COMPLEX/PFM	7.7	2.5
SHORTZ	6.9	2.1
IMPACT	10.0	0.5

(*Impact was run only on selected portions of the WESTVACO data.)

Clearly, gross overprediction is the rule for all of the models other than RTDM on the industrial SO_2 data, and overprediction still appears in the passive tracer data, except for IMPACT, which underpredicted. Curiously, IMPACT was a serious overpredictor in the SO_2 evaluation.

V. CONCLUSIONS

The AMS members of the Committee feel that the following are the key findings of the evaluation. These points are not a synthesis of the findings of the reviewers, although each of them concurs in several if not all of the conclusions.

- The models and the data sets against which they were compared provide a very limited representation of the full variety and range of the flow patterns and diffusion situations actually occurring in complex terrain. The evaluation has provided limited data on predicted and observed concentrations that might be expected on nearby slopes and crests, nothing more. Problems involving internal valley circulations, diffusion in meandering valleys, slope flow, and terrain more distant than the immediate obstacles cannot be assessed with the models discussed in this study.

- The voluminous performance measures developed in this study, like those in the rural and urban reviews before it, are difficult to digest. Both the reviewers and the Committee would have been happier if the comparison against the data had provided more information about what the models were doing in a limited number of specific situations.

- It is highly unlikely that complex terrain problems can ever be treated intelligently with a set of "cookbook" models. Flow and diffusion in complex terrain is, and will continue to be, complex.

REVIEW OF COMPLEX TERRAIN MODELS

Prepared for the AMS-EPA Steering Committee

by

Jason K.S. Ching *
Meteorology and Assessment Division
Atmospheric Sciences Research Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711

April 1985

---

* On assignment from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce.

ABSTRACT

A two stage review of eight different complex terrain dispersion models was performed. First, the scientific bases for these models, both separately and as an aggregate, were examined. Second, the operational performance of these models against two sets of complex terrain model evaluation data bases as prepared by TRC Environmental Consultants, Inc. were reviewed. These models: COMPLEX I, COMPLEX II, COMPLEX PFM, 4141, PLUME5, RTDM, SHORTZ and IMPACT and references to the documentation and information of each model reviewed are found in the reference list of the TRC draft report "Evaluation of Complex Terrain Air Quality Models", Wackter and Londergan (1984). The first seven models listed are Gaussian type models; the eighth, IMPACT, is a deterministic finite difference numerical grid diffusion model.

The approach taken to parameterize plume rise, dispersion, transport, terrain adjustment, and mixed layer heights by each model was examined. Each model's treatment of terrain features and the stability dependency of flow over terrain and its potential importance to complex terrain modeling was studied in greatest detail. It was found that terrain adjustment and transport was treated quite differently by all the models in this set; in combination with other model components, it was concluded that large differences in model predictions is expected. The incorporation of potential flow concepts and Froude number scaling and the use of a critical dividing streamline height by some models is clearly an important improvement.

The performance analyses, evaluation and conclusions drawn by TRC as reported in Wackter and Londergan was useful, but did not provide sensitivity diagnostics. Additional model comparisons are included to compliment the TRC study in an effort to compare the different model predictions relative to the methodology adopted by individual models in handling the more important and sensitive model components. It was shown that model prediction differences, and errors can be as much a function of inadequate model methodologies as well as in the formulation of the basic model components.

TECHNICAL EVALUATION

Introduction

Air quality simulation models used for assessing the air quality impact of point sources are based on solving the standard diffusion equation. Typically, model approaches include numerical simulations on a prescribed grid or the adoption of the Gaussian solutions with ad hoc approaches to specific circumstances. The first approach permits a high level of realism, but computational requirements and model complexities increase with the level of realism and accuracy required. Gaussian models are based on simplistic solutions to the diffusion equation. It is a highly parameterized scheme based on a few empirical studies of simple flows over uncomplicated surface conditions, but it is clear that most of the applications quickly overextend the physical and empirical bases. [Improvements in the state-of-the-art Gaussian models are possible with (1) improvements in the parameterizations by extending the various parametric theoretical or empirical bases and by (2) more general and consistent ad hoc engineering modeling approaches to the input data, to boundary conditions and to handling of unique conditions.] The Gaussian modeling approach is highly practical and relatively inexpensive to use as an assessment tool. Unfortunately, in most instances, the initial application of a model (even a sophisticated one) to a specific problem will yield potentially large errors, much of which may arise out of inconsistences between the theoretical limitation and the unique, ad hoc approaches specific to each model. A recent article by M. Smith, (1984) provides a synthesis of seven independent reviewers' evaluations of Rural Air Quality Simulation Models. In that cooperative study between the U.S. EPA and the American Meteorological Society (AMS), 10 models were studied and evaluated against a common data base, much as was done in the current study. In particular, COMPLEX I and II, PLUME5 and 4141, currently being evaluated, were also evaluated in that study. Additionally, a similar evaluation study was conducted for the EPA/AMS on a comparable set of Urban Air Quality Simulation Models. In general, both studies reveal the scientific bases of these models to be out of date. Recent improvements in our knowledge of the structure and evaluation of the planetary boundary layer (PBL), especially during the convective and also during the nocturnal periods, have not been incorporated into these models for example. Further, stagnation conditions cannot be handled by these Gaussian schemes, and the stability and dispersion parameterization classified according to Pasquill-Gifford-Turner developed from surface data may not be appropriate for tall stacks. I fully support Smith's (1984) viewpoint and the bases for his summary, including a statement regarding inherent uncertainties in both predictive schemes and in the inadequacies of the input and the receptor data bases.

The technical and scientific bases of the Gaussian models in Smith's 1984 article have not been appreciably improved in the present set of models. Therefore, relatively speaking, the scientific bases are considerably more limiting for model application in complex terrain where each of the processes modeled is affected in some way by the underlying terrain. There are recent theoretical concepts that provide the first real break-throughs in extending the Gaussian modeling approach from flat to complex terrain. First, a basis for neutral flow over simple two and three dimensional obstacles has been postulated based on potential flow theory and fluid modeling tank experiments. Second, in stable conditions, a critical dividing streamline height can be computed based on Froude number scaling below which the flow cannot surmount the obstacle. This is examined in greater detail below.

All the models in this study compute one, three and twenty-four hour short term averages from which long term predictions such as annual averages can be calculated. Additionally, annual averages are also calculable directly by some models. Each model handles multiple sources and large numbers of receptors. Input meteorology, emissions and terrain data were treated somewhat differently between each model. Default procedures for handling special limitations such as missing data, stagnation conditions, etc., varied between models. No two models were exactly alike. For one, each Gaussian model adaptation to complex topography was different. Subsequently, these models all varied in their predictive skills against the two common data bases used in the TRC evaluation. The most sophisticated models recognized Froude number scaling and incorporated critical height for flow separation (RTDM and COMPLEX/PFM). Bounds were established which limit reflection to contribute no more to surface concentration than would satisfy the constraint of the 2nd law of thermodynamics wherein, the surface concentration cannot increase with downwind distances past the first maximum (RTDM). Plume centerline is adjusted for terrain differently for all models. Specific model components for each model is discussed below.

Specific Model Components

Some of the major processes in complex terrain to be modeled include plume rise, horizontal and vertical dispersion, transport, transformation and deposition which are all influenced by the variations in the underlying terrain. Table 1 lists some of the more important aspects of each of these processes and the extent to which they are included in the eight models. In several instances, the treatment of some particular aspect was either not ascertainable or ambiguous from the available documentation, and is so indicated by an X. The available model documentation varied considerably in the degree of specificity for different model components.

Plume Rise

Plume rise is an important dispersion process for buoyant point sources. Effluents may experience downwashing beyond the stack but typically rise to a final equilibrium height (H_a) where no excess buoyancy remains. A finite distance is traversed during plume rise and environmental air is entrained into the plume during this transition period. Plume growth after final plume rise is dependent on the magnitude of the turbulent intensity of the environment. Plumes during this transition phase don't impact ground level concentration unless downwind receptors are located above stack top, as might be the case for complex terrain. The theory and formulations for transition and final plume rise by Briggs (1975 and 1984) are generally the accepted standards and with the exception of SHORTZ and PLUME5 have been adapted by all the models in this set. The presence of

Table 1. Complex Terrain Model Component Features

Feature	COMPLEX I	COMPLEX II	COMPLEX PFM	4141	Plume5	RTDM	SHORTZ	IMPACT
Plume: Transitional Plume Rise	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Partial P.R. Penetration	X	X	(X)	X	No	Yes	No	Yes
Momentum Induced	Yes	Yes	Yes	X (CRSTER) Yes (MPTER)	X	No	No	X
Stack Tip Downwash	Yes	Yes	Yes	X (CRSTER) Yes (MPTER)	X	Yes	Yes	X
Dispersion: Bivariate Gaussian	No	Yes	Yes	Yes	Yes	Yes	Yes	N/A
Buoyancy Induced Dispersion	3.5	3.5	3.5	4.3	3.5	√10	4.3	Yes
Enhanced by Wind Shear	No	No	No	No	Yes	Yes	Yes	N/A
Fumigation	No	No	No	No	No	No	No	X
Dispersion Parameters from observations	No	No	No	No	No	Yes	Yes	No
Transport: Critical height/flow	No	No	Yes	No	No	Yes	No	X
Separation	No	No	No	No	No	Yes	No	X
Partial Reflection	No	No	No	No	No	Yes	Yes	No
Wind profile power law from obs.								
Transport wind to a) stack top	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
b) final P.R.	No	No	No	No	No	Yes	Yes	Yes
Deposition	No	No	No	No	No	No	X	No
Terrain Adjustment	Yes	Yes	Yes	Yes	Yes	Yes	Partial	Yes
Transformation	No	No	No	No	Yes	No	X	Yes

ground based and elevated stable layers, will strongly control plume rise and dispersion. Portions of plumes that penetrate elevated inversions, Z_i, will not impact the near ground level unless receptors are located at or above this level. Thus, the extent to which plumes penetrate elevated stable layers greatly influences the subsequent ground level concentration. In effect, this process acts as a source depletion term and must be modeled accurately. Most Gaussian plume models have adopted an all-or-nothing approach whereby only those plumes whose plume rise is less than the mixed layer height can impact the surface. Such models are burdened with the need to model H_a and Z_i very accurately in order to correctly determine the sign of $H_a - Z_i$, but both parameters are known to display large variability and accurate modeling of these parameters and their difference is apt to be unsuccessful. This methodology is so critical that extremely large overpredictions or underpredictions are expected.

COMPLEX/PFM (hereinafter PFM), RTDM, and IMPACT are the only models in this set that address this modeling of plume penetration issue by incorporating a layered plume rise calculation scheme. The other schemes are apt to exhibit very large predicted variances due to the all-or-nothing plume rise into elevated stable layer methodologies. PFM calculates plume rise in layers determined from hourly wind and temperature profiles interpolated from the twice daily radiosondes. The reduction of buoyancy flux is calculated layer by layer until the final plume rise. In this fashion, PFM accounts for elevated inversions and vertical wind shears. However, this calculations is still an all-or-nothing calculation within any of these stable layers aloft. RTDM and IMPACT calculate the fractional amount of plume imbedded in the elevated stable layers according to the Briggs (1975) rectangular plume model. Documentation for PLUME5 was unclear on this point. Perhaps the most acceptable means of modeling plume penetration into elevated stable layers are to be found in Briggs, (1984) and Weil and Brower (1982). Briggs (1984) assumes the mixed layer stability near the inversion base to be the same as that of the thermal gradient in the elevated stable layer. Weil and Brower (1982) suggest a more realistic value in the mixed layer. None of the models reviewed utilize either approach.

Generally, plume rise is predominantly buoyancy driven. However, in some instances, it would be more accurate and consistent to include the contribution by momentum. Currently, RTDM and SHORTZ exclude this feature; it was indeterminate for 4141, PLUME5 and IMPACT. The inclusion of stack tip downwash was generally available as an option for most of these models. These last two features are not expected to be critical features of plume rise as it impacts complex terrain model predictions unless the emissions are predominantly non-buoyant.

Dispersion

After initial buoyancy-induced plume spread, Gaussian plume models utilize bivariate parameters for plume spread in the vertical and lateral directions. This spread is prescribed to be a function of atmospheric stability and is a function of turbulent intensity and travel distance. Default values rely on the use of empirical sigma curves according to stability classes known as attributed to Pasquill, Gifford and Turner. The experimental bases for these functions are for relatively flat terrain, near surface, nonbuoyant plumes and the general applicability of these curves to complex terrain and to tall stack, buoyant plumes is yet to be established. With this in mind, the operational complex terrain model that appropriately incorporates on-site turbulence data (intensity or sigma theta, phi) for hourly averages is preferred by this reviewer. Only RTDM and SHORTZ provide a user option to compute σ_y, σ_z from input, on-site, turbulence data. In the meantime, studies to derive generalized dispersion parameterization schemes applicable to complex terrain are highly recommended. Currently, several major field studies programs are underway in Complex Terrain which should provide for improvement in dispersion parameterization (ASCOT, see Gudiksen and Dickerson, 1983; EPA Complex Terrain Model Development Study, Lavery et al., 1982, 1983, 1984; and Plume Model Validation Study, Hilst, 1978).

With the exception of COMPLEX I all the Gaussian models of plume dispersion in this set were bivariate. COMPLEX I considered the lateral spread a constant of 22 1/2°, thus independent of stability. IMPACT computes dispersion using eddy diffusivities rather than sigma specifications. In the current version, IMPACT calls a subroutine (DEPICT) which computes K_z=K u $\sigma_\phi \ell$ where u=wind speed, σ_ϕ is standard deviation of vertial wind vane fluctuation expressed in terms of stability classes and ℓ is a turbulent length scale, also dependent on stability classes as well as altitude. These are provided in lookup tables. With the exception of RTDM and SHORTZ, the dispersion parameters for the remaining Gaussian models consisted of PGT or Briggs curves where the stability classes for COMPLEX I, II, PFM, and 4141 were determined using standard Turner method, PLUME5 utilizes σ_ϕ data, RTDM use σ_ϕ data and IMPACT uses temperature gradient input data. RTDM and SHORTZ use input vertical and horizontal turbulence intensity data to compute the dispersion coefficients. Note that a stability class change by one corresponds to a change in dispersion by about a factor varying around two or more; thus, the determination of stability classes is very critical. An interesting comparison between the diffusivity schemes of IMPACT and the corresponding PGT dispersion curves was performed by RADIAN and discussed in their model documentation. The analyses suggest large differences in the diffusivity used by IMPACT versus those of the Gaussian model for the various stability classes. This survey would suggest that the methodology for determining hourly dispersion coefficients differs among most of the models in this set, and it would be of interest to conduct a detailed comparative analyses of the dispersion parameters calculated hourly for each of the models. Much of the differences between model prediction might arise out of differences in these dispersion calculations. Recent and significant advances in dispersion theory (Lamb, 1978, 1979, 1981, and Willis and Deardorff, 1978, 1983) and observational studies of tall, buoyant stack dispersion applicable to convective mixed layers (Weil, 1979) provide the bases for greatly improved dispersion parameterization schemes such as proposed by Weil (1979) and Weil and Brower (1982 and 1983). Such formalisms supercede the artificial restrictions of relating dispersion to distinct stability classes.

With the exception of IMPACT, each model included an adjustment to plume spread arising out of buoyancy-induced entrainment of environmental air while undergoing plume rise. Each considers this spread as $\Delta h/\alpha$ where α is a proportionally constant and Δh is plume rise during transition. Table 1 lists values of α for each of the models. Differences are as large as 36%, yet relatively insignificant with respect to uncertainties associated with other parameters. An additional enhancement of plume spread occurs as a result of vertical wind shear acting over plume rise. RTDM, SHORTZ and PLUME5 include this enhancement which depends on wind shear through plume rise and distance traveled. Baroclinic condition, nocturnal wind shear all quite prevalent climatically suggest that this feature attains considerable importance in complex terrain for conditions where receptors are at altitudes in the vicinity of effective stack heights. Pasquill's suggested relationship of $0.17 \cdot \Delta\theta \cdot x$ can be comparable to dispersion induced by atmospheric turbulence, especially for the more stable dispersion classes.

There is evidence to suggest that fumigation after sunrise and incidences of high ground level pollutant concentration are correlated, both in simple and in complex terrain (Frank, Blagun and Slater, 1981), yet this process is ignored by all the Gaussian models. It would be of interest to determine whether IMPACT simulates this process; the current documentation does not permit this examination. The TRC model evaluation did not provide analyses sufficiently detailed to determine how critical the process is in complex terrain, but it is suggested that both data sets could provide a limited basis for such an analysis.

Transport

Recent theories by Hunt and Mulhearn (1973) Hunt, Puttock and Snyder, (1979), and Hunt and Snyder (1980) provide a bases for dispersion modeling applicable to non-flat surfaces. Utilization of potential flow theory permits the generation of streamlines over a variety of surfaces for neutral to slightly stable conditions. The adjustment of the flow field over the obstacles causes distortion in plume spread due to compression and expansion of the streamlines. For flows that are stable or for large obstacle sizes, a critical dividing streamline height, HC, exists above the obstacle base for which plumes whose center lines are below HC do not flow over the obstacles (i.e., ground impaction), while plumes above HC rise over the impediment. This critical height reduces to zero in neutral to unstable stability and the plume then flows over the obstacle at all elevations. There is no comparable theoretical or physical bases for the unstable or convective boundary layer. One intuitively expects some undulation of the plume relative to terrain; this undulation can be as large as the underlying terrain or only slightly distorted. A half-height correction is recommended by the Specialist Conference on EPA modeling guidelines (Chicago, 1977) as a suitable and conservative terrain adjustment procedure for screening purposes, until the advent of more fundamental empirical bases. Airborne lidar studies (Uthe, 1984) show examples of plumes over terrain which depict undulations almost parallel to the underlying terrain.

Given this background, we see that only RTDM and PFM utilize potential flow approximation and dividing streamline concept. Without a doubt the most physically consistent and sophisticated Gaussian model in this set is PFM. It is by far the most detailed and complete in introducing two and three dimensional terrain features and some engineering approaches to incorporate oblique wind direction angle of attack to the major terrain axis. Further, the relationship for plume spread σ_y and σ_z over flat terrain is calculated as modified by potential flow theory. PFM, however, uses a rather awkward scheme to compute concentrations; under stable conditions, concentrations for receptors below the height of the dividing streamline HC are calculated using COMPLEX I, the monovariate dispersion model and the concentration above HC utilizes the PFM modified flow fields. PFM defers to COMPLEX II for its concentration computation under unstable conditions. PFM doesn't provide for partial reflections to limit concentrations from increasing with distance as might be possible due to influence of large terrain slope on multiple reflection calculation scheme.

The RTDM recognizes Froude number scaling and the dividing streamline concept. This model surveys the gridded terrain contours for each 10 degree sector and tabulates the distance from the source to the successive predetermined contour intervals, stopping at the highest point within this sector. RTDM ignores receptors at successive terrain features farther downwind, a somewhat artificial restriction, but consistent for screening considerations. Note that RTDM makes no adjustment to the sigma curves over terrain as PFM does. RTDM is the only Gaussian model in this set to limit concentrations from increasing with downwind distance to avoid violating the principle of the 2nd law of thermodynamics from a multiple reflection calculation scheme applied to a steep terrain feature.

RTDM and SHORTZ are the only models to extract their wind profile power laws from on site data, and along with the IMPACT model consider the transport wind the average between stack top and final plume rise. (where is eq. 17 in Plume 5 documentation?)

None of the models include surface uptake by dry deposition, (i.e., source depletion) and capabilities for photochemistry was limited to PLUME5 and IMPACT.

Terrain adjustment

Believing that methodology for treatment of plume and flow adjustment to underlying terrain is critical in complex terrain models, this aspect was examined in some detail. I thought it useful to provide the details of my comparative analyses (all figures in this review are taken from the documentation provided). I concluded that none of the models (except for COMPLEX I and II) utilize the same modeling procedures for adjustment of plume height over terrain. It is suggested that large variance in prediction may be caused inadvertently by a variety of rather artificial modeling procedures. This will become evident when comparing model output against the Westvaco data set as discussed later. The clarity and quality of the model procedures for terrain adjustment varied widely between model documentation. COMPLEX I and II handle terrain adjustment as indicated in Figure 1a. Thus,

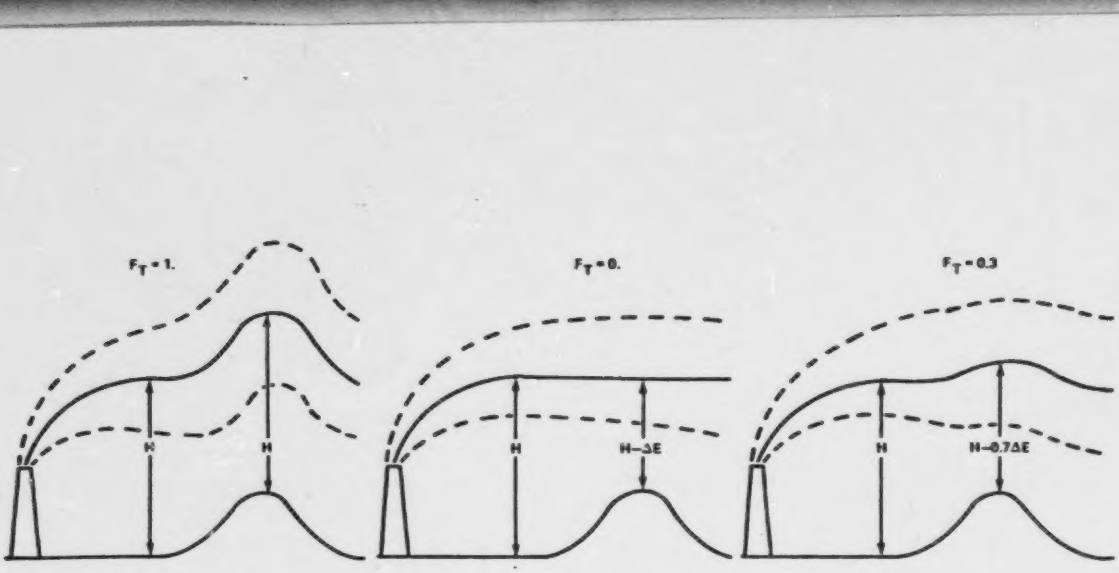

F_T = 1.
H
H
F_T = 0.
H
H−ΔE
F_T = 0.3
H
H−0.7ΔE

Figure 1a. COMPLEX I, II (PFM), 4141.

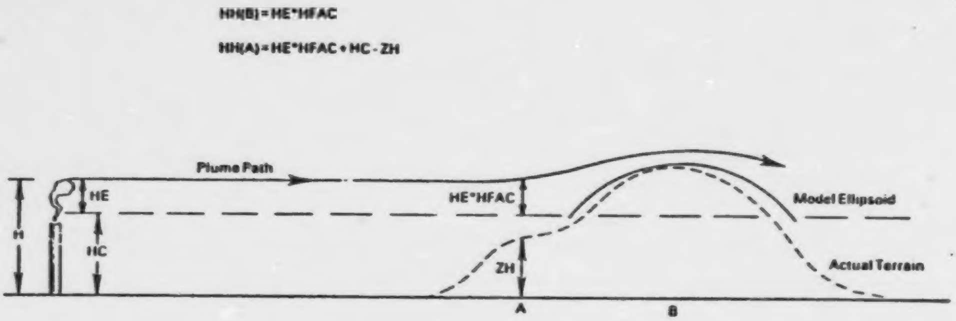

HH(B) = HE*HFAC
HH(A) = HE*HFAC + HC - ZH
Plume Path
HE
H
HC
HE*HFAC
Model Ellipsoid
ZH
Actual Terrain
A
B

Figure 1b. (COMPLEX PFM).

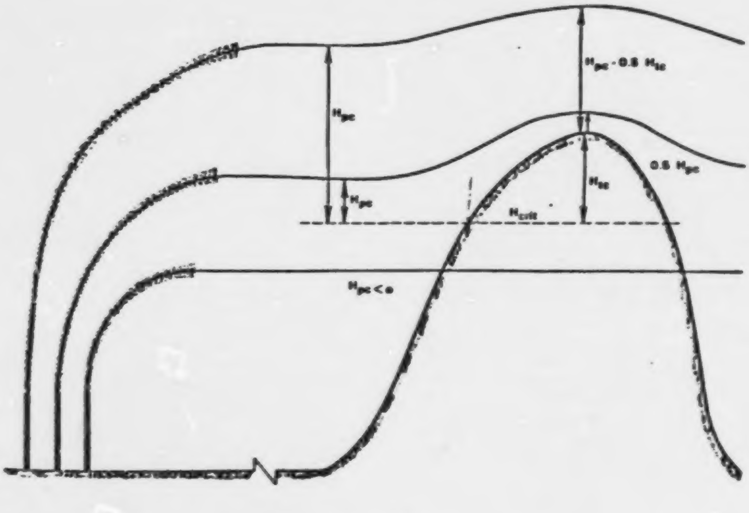

H_pc
H_pc - 0.5 H_tc
0.5 H_pc
H_tc
H_pc
H_crit
H_pc < 0

Figure 2. (RTDM).

$$H_a = H-(1-F_T)\Delta E$$

where H_a is the adjusted effective plume height

H is the equilibrium plume rise above flat terrain
ΔE is $ZR-ZS$ where
ZR is receptor height and
ZS is elevation of base of source
F_T is terrain adjustment factor

For unstable to neutral conditions, F_T is set to 0.5; and is zero for stable flows. Terrain adjustments are limited to receptors lower than the top of the lowest stacks modeled. Thus,

$$H_a = H-0.5\ \Delta E \text{ unstable}$$
$$H_a = H-\Delta E \text{ stable}$$

There is no provision for calculating the height of the dividing streamline, or HC in these two models.

The terrain factor is zero for stable flows; thus the plume is horizontal and the plume can impact the surface for all $\Delta E \geq H$. In this case, COMPLEX I and II perform VALLEY-like impingement calculations. These two models do not recognize a dividing streamline height; plume rise above the mixed layer height is ignored.

In PFM, for stability classes D, E and F, for $H_a > HC$

$$H_a = (H-HC)\ HFAC \text{ for } Z_R > HC$$

and

$$H_a = (H-HC)\ HFAC + (HC-ZR) \text{ for } ZR < HC$$

HFAC is determined from application of potential flow theory and the plume path shown in Figure 1b to be along a streamline at plume rise, H. In this situation, concentration for receptors at heights less than HC are zero.

$$H_a = \frac{H_a}{2} \left(\frac{\Delta E - 0.5 H_a}{0.9 Z_B - 0.5 H_a} \right) + 1)$$

This approach is quite different in form from the previous models (there is no simple form for F_T for example), and depends on Z_B, the height of the base of the elevated inversion. This elaborate technique was not justified on any physical principle.

RTDM recognizes a critical height HC for which only the air above this level can flow over the terrain; plumes below will impinge onto the terrain. For neutral to unstable conditions, HC=0. Using HC as the reference level, RTDM then utilizes a plume path coefficient, C, which is exactly the same as F_T with default value of 0.5 for all stabilities (as was used in the TRC model performance). Flow immediately above HC have little but sufficient energy to crest the terrain, so

$$H_a = F_T \, H_{pc}$$

where H_{pc} = height of plume above HC

($F_T = 0$ if $H_{pc} < 0$, plume below HC implies terrain impaction)

For plume height at or above terrain crest,

$H_a = H_{pc} - (1-F_T)H_{Tc}$ where

H_{Tc} is height of terrain above H_c.

This methodology is illustrated in Figure 2.

For plumes below HC, the plume centerline is not adjusted for terrain and COMPLEX I is invoked to perform the impingement calculations, which is known to be conservative prediction.

When HC is zero, as in neutral to unstable conditions, it can be shown that $HFAC = 1 - \frac{0.5\Delta E}{H}$ for $F_T = 1/2$ as in COMPLEX I and II. For these flows, PFM utilizes COMPLEX II to compute the concentration fields, after plume rise and mixed layer depths have already been computed using the PFM approach.

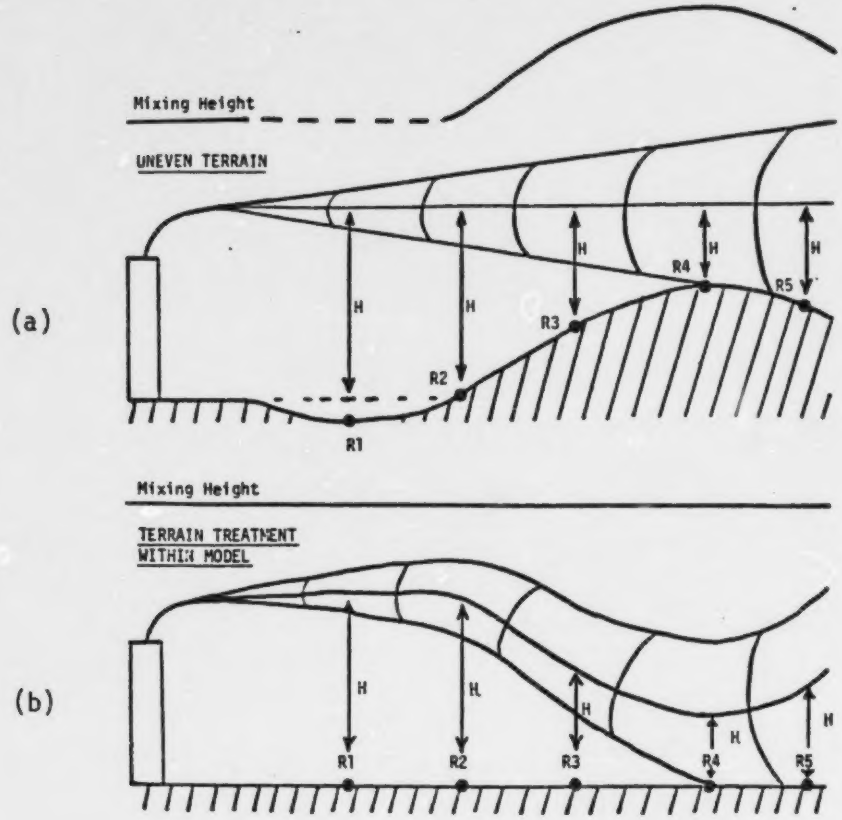
Model 4141 also shown in Figure 1a utilizes a modified CRSTER approach, where,

$H_a = H - (1-F_T)\Delta E$
$F_T = 0.5$ for A, B, C and D stability and
$= 0.25$ for E, F stability.

This differs from COMPLEX I and II in the stable classes and is less conservative. ΔE is limited to positive values, i.e., receptors are assumed to be no lower than stack base.

If H_a is less than zero, the effective stack height is set to zero and the plume impacts the surface. Moreover, terrain adjustment is limited to those receptors at heights less than the top of the lowest stack considered. This stipulation is the same for COMPLEX I and II.

PLUME5 (see Figure 3) considers plume rise relative to the location of the stable layer. When H_a above the top of the stable layer $\chi_0 = 0$ for receptors below Z_i and also for receptors outside a stable layer for which H_a is imbedded. In the case of H_a below elevated stable layers over terrain, $F_T=0$ for receptors located below $H_a/2$; and $H_a = H-\Delta E$. (i.e., plume impaction possible). For receptors above $H_a/2$ the relative height of the terrain relative to plume rise is


Mixing Height
UNEVEN TERRAIN
(a)
H
R1
R2
R3
R4
R5
Mixing Height
TERRAIN TREATMENT WITHIN MODEL
(b)
H
R1
R2
R3
R4
R5

Note: R1-R5 are receptor points at 5 ring distances.

Figure 3. (PLUME5).

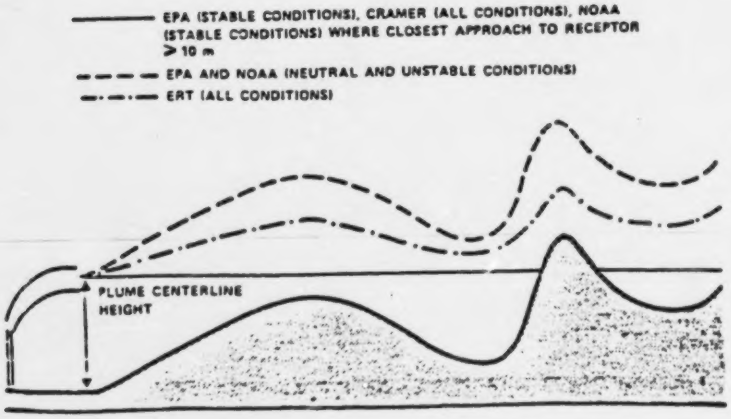

EPA (STABLE CONDITIONS), CRAMER (ALL CONDITIONS), NOAA (STABLE CONDITIONS) WHERE CLOSEST APPROACH TO RECEPTOR > 10 m
EPA AND NOAA (NEUTRAL AND UNSTABLE CONDITIONS)
ERT (ALL CONDITIONS)
PLUME CENTERLINE HEIGHT

Figure 4. (IMPACT).

When the flow is neutral to unstable $H_{pc} = H$

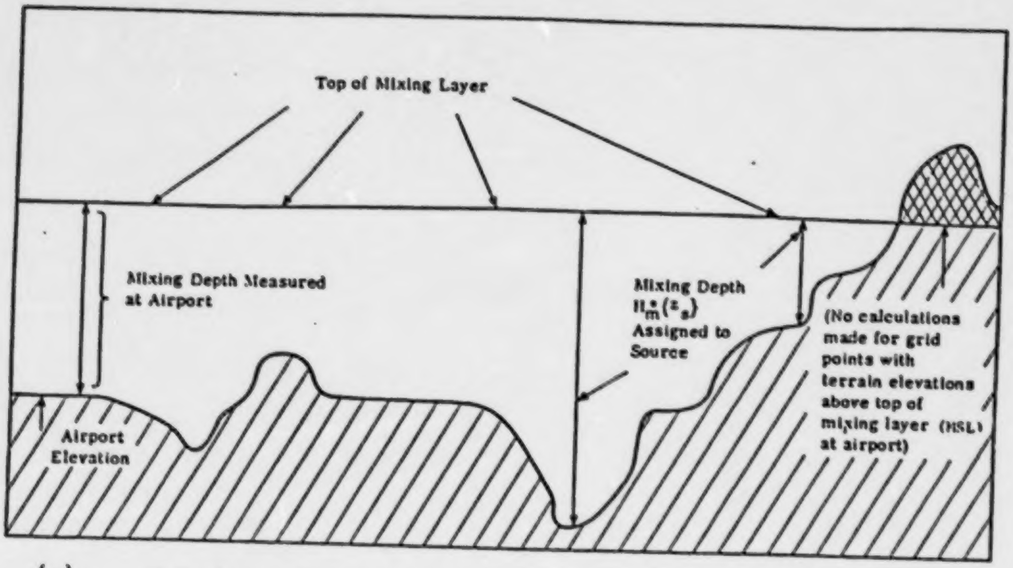
and $H_a = F_T H$ for $H < \Delta E$

$H_a = H - (1-F_T)\Delta E$ for $H > \Delta E$

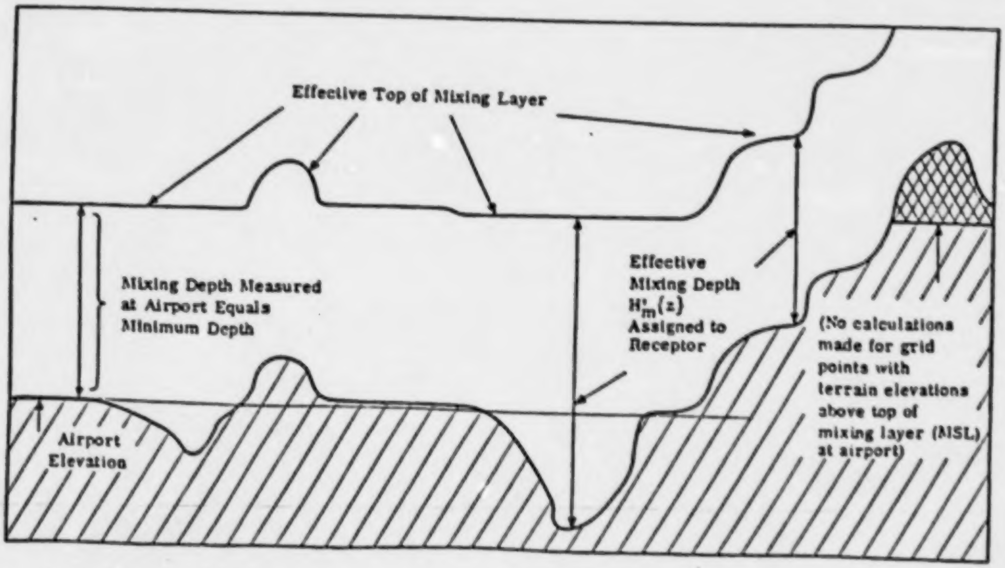
Then, since $F_T = 0.5$ this formulation is the same as for COMPLEX I and II for $H > \Delta E$ for those stability groups, but without the limitation on a receptor to be lower than the top of the lowest stack. Appropriate adjustments are made to Z_i to avoid inconsistences.

The method utilized in SHORTZ is illustrated in Figure 5a and b. SHORTZ restricts ground level concentration calculations to receptors below H_m where H_m is the actual mixed layer height above the source and is not modified by terrain (Figure 5a). Thereafter, for computational purposes, an effective mixed layer height (Figure 5b) is utilized, restricted to be no less than H_m. Plumes above H_m do not contribute to ground level concentrations. Plume centerlines below H_m are presumed to be unaltered by presence of terrain (i.e., $F_T = 0$). The SHORTZ approach is thus an ad hoc engineering method without a sound physical basis. Statistics that are heavily weighted by receptors generally located above H_m will be underpredicted. Receptors below H_m on an average will overpredict. Thus, model performance statistics are complicated by opposing tendencies for overprediction due to lack of terrain adjustment to plumes, and the assignment of zero concentrations to receptors located above the mixed layer, the latter expected to be more pronounced during the night under stable conditions when H_m is likely to be low. Their use of the term "surface mixing layer" is a bit archaic and can be misleading since they really intend the "mixed layer". (Section 2.1.1.1, alluding to a definition of mixing depth by use of turbulent intensity rather than by thermal stratification alone, is missing in their documentation).

IMPACT is unique among this set. It is a deterministic finite difference numerical grid model which predicts the concentration fields in three dimensions based on the mass conservation laws for species. The flow field


Top of Mixing Layer
Mixing Depth Measured at Airport
Mixing Depth $H_m^*(z_s)$ Assigned to Source
(No calculations made for grid points with terrain elevations above top of mixing layer (MSL) at airport)
Airport Elevation

(a) Mixing depth $H_m^*(z_s)$ used to determine whether the stabilized plume is contained within the surface mixing layer.


Effective Top of Mixing Layer
Mixing Depth Measured at Airport Equals Minimum Depth
Effective Mixing Depth $H_m'(z)$ Assigned to Receptor
(No calculations made for grid points with terrain elevations above top of mixing layer (MSL) at airport)
Airport Elevation

(b) Effective mixing depth $H_m'(z)$ assigned to receptors for the concentration calculations.

Figure 5. (SHORTZ).

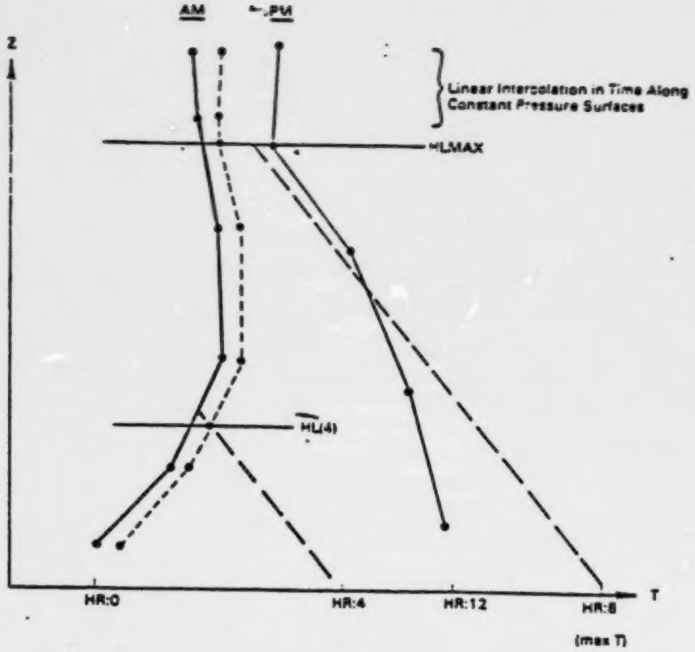
is non divergent throughout and terrain following at the surfaces. Dispersion from point sources into the flow field is accomplished by means of diffusivity prescribed as a function of stability. The accuracy is determined to some extent by the grid resolution and must be chosen carefully to balance both practical computational requirements and accuracy. Terrain considerations are thus handled implicitly as part of the numerical model rather than explicitly as in the Gaussian models.

Mixed layer height (H_m)

The mixed layer height is an important dispersion parameter in Gaussian models. Certainly, this term appears explicitly as a reflection condition. Its criticality is most pronounced, however, with respect to criteria for partial plume penetration. H_m remains a reflecting surface for plumes until plume rise is larger relative to H_m. Then H_m may be modeled as an insulating surface. In complex terrain, receptors may lie above, at or below H_m and thus the models predicted concentration will be subject to the accuracy of predicting H_m and H_a as well as the modeling strategy. It can be shown that enormous prediction errors are possible due to improper modeling strategy using the Gaussian model. Numerous mixed layer growth theories and parameterization approaches now exist, especially for the convective boundary layer. None of these models apply such approaches, however.

Typically, H_m exhibits large diurnal changes which are modulated by synoptic scale conditions and by season and geography. In the present case of complex terrain modeling, topography will certainly influence the spatial distribution of H_m, as well as H_a as discussed earlier. For this latter reason, the accuracy of Gaussian models is truly sensitive, even critically dependent on model methodology for plume rise H, or H_a relative to H_m.

COMPLEX I and II are similar. Mixed layer heights are preprocessed from twice daily radiosondes soundings and linearly interpolated for hourly values (see Figure 6). H_m is linearly interpolated from the previous day's maximum value to the present day maximum unless the surface is stable at



Sample construction of temperature profile for fourth hour after A.M. sounding.

Figure 6. COMPLEX (I, II, PFM).

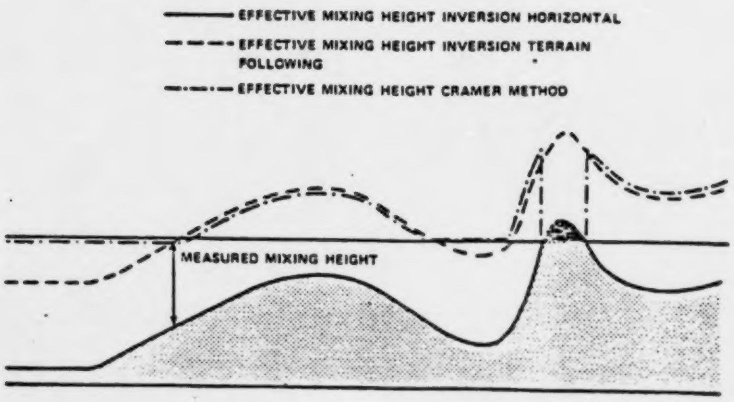

EFFECTIVE MIXING HEIGHT INVERSION HORIZONTAL
EFFECTIVE MIXING HEIGHT INVERSION TERRAIN FOLLOWING
EFFECTIVE MIXING HEIGHT CRAMER METHOD
MEASURED MIXING HEIGHT

Figure 7. (IMPACT).

sunrise at which time the interpolation proceeds linearly from about sunrise using the morning minimum mixed layer values and the standard Holzworth (1972) technique. A plume with final rise above H_m is considered insulated from the surface. This interpolation scheme for H_m is rather crude and lacks temporal resolution to deal with the rapid rise in the morning transition period. Plume rise, discussed earlier, has a 1/2 height terrain adjustment and thus, if H_m is level, the assumption of insularity will cause the average predicted concentration at receptors above H_m to be relatively low, and those at or just below to be maximum (i.e. near plume centerline). Thus large variations are expected using those models, and especially if receptors are distributed such that they are above the stable inversions at night and below during the day. Note also that terrain factors are limited to those receptors below stack top. Since plume rise relative to mixed layer heights is so critical, model predictions for a hypothetical set of receptors lower than stack top will be quite different from those receptors above stack top. This restriction seems unnecessary, arbitrary and defeating and should be considered for removal.

The PFM methodology for determining the temporal variation of H_m is considered the superior approach in this set of Gaussian models that utilizes twice daily soundings. It is an improved Benkley-Schulman (1979) approach providing better vertical resolution of the temperature advection. Unfortunately, no allowance is made to alter mixed layer height as a function of topography, even though its terrain adjustment procedure was probably the most sophisticated and soundly based theoretically. Thus, it too, as with other similar models is expected to suffer large predictive variances. Layered plume rise certainly provides improvements over the COMPLEX models.

Model 4141 (see Figure 8) utilizes procedures for the MPTER model in developing hourly mixing heights. This procedure is identical to COMPLEX I and II.

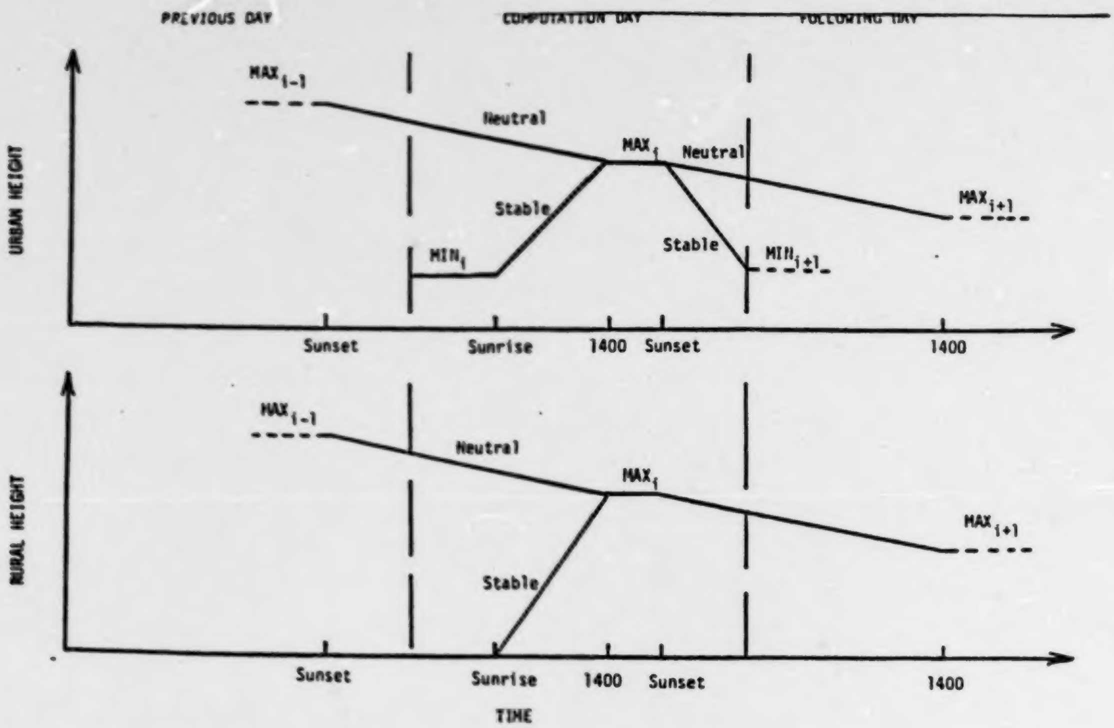
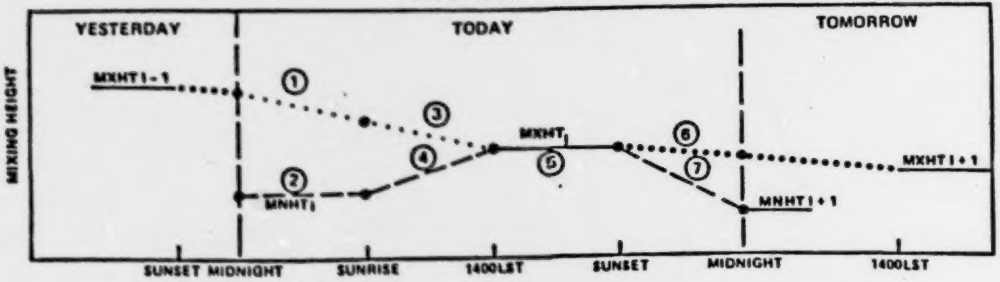

PREVIOUS DAY
COMPUTATION DAY
FOLLOWING DAY
URBAN HEIGHT
MAX_{i-1}
Neutral
MAX_i
Neutral
Stable
MIN_i
Stable
MIN_{i+1}
MAX_{i+1}
Sunset
Sunrise
1400
Sunset
1400
RURAL HEIGHT
MAX_{i-1}
Neutral
MAX_i
MAX_{i+1}
Stable
Sunset
Sunrise
1400
Sunset
1400
TIME

Figure 8. Mixing height as used by 4141.

(a) URBAN


YESTERDAY
TODAY
TOMORROW
MIXING HEIGHT
MXHT I - 1
①
②
③
④
⑤
⑥
⑦
MXHT I
MNHT I
MXHT I + 1
MNHT I + 1
SUNSET
MIDNIGHT
SUNRISE
1400LST
SUNSET
MIDNIGHT
1400LST

(b) RURAL 2

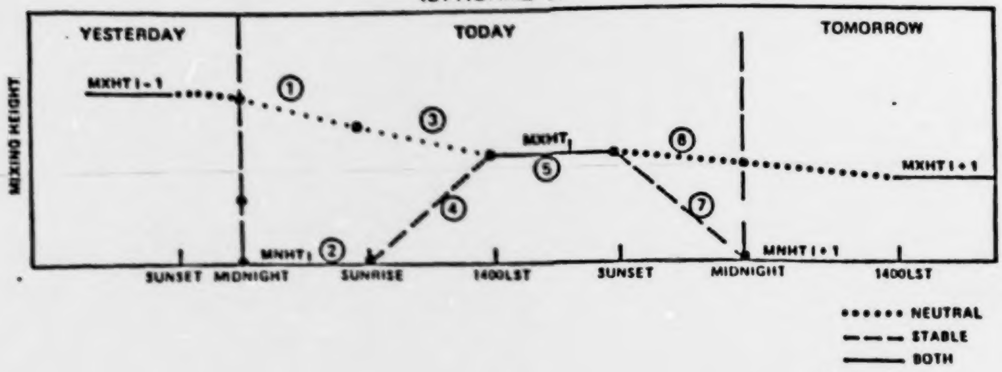

YESTERDAY
TODAY
TOMORROW
MIXING HEIGHT
MXHT I - 1
①
②
③
④
⑤
⑥
⑦
MXHT I
MNHT I
MXHT I + 1
MNHT I + 1
SUNSET
MIDNIGHT
SUNRISE
1400LST
SUNSET
MIDNIGHT
1400LST
NEUTRAL
STABLE
BOTH

Figure 9. MIXING HEIGHT ALGORITHMS USED IN PLUME5.

PLUME5 (see Figure 9) utilizes a similar procedure for determining mixed layer heights as COMPLEX I and II with the following difference. If the surface layer is stable between sunset to midnight and from midnight to sunrise, the model uses a minimum value of H_m determined from the morning sounding, otherwise h is obtained by interpolation between the previous and current days maximum H_m. If the hour before sunrise is stable, the mixed layer is obtained by linear interpolation between the min H_m of the morning soundings and the max H_m of the afternoon sounding. This is a significant difference since H_m doesn't drop to a minimum value for stable conditions at night in COMPLEX I and II and 4141, but it does for PLUME5. The ramifications on concentration predictions will probably cause large differences in model outcome between those two approaches as discussed above.

RTDM doesn't specify any modification to procedures used in the CRSTER model for computing its mixed layer heights. In this regard RTDM methodology is identical to COMPLEX I, II and 4141. But terrain adjustments are considerably different as discussed earlier.

SHORTZ model strongly recommends the option of user input mixed layer data. However, it defaults to CRSTER-like calculation but substitutes a 2.5 times the significant roughness element height, Z_0 for minimum H_m in stable conditions. This procedure is totally arbitrary however. For one, the user has no guidance for a proper choice of appropriate Z_0. Once determined, the H_m value is used to eliminate receptors from consideration in concentration prediction for the hour when the receptors are above H_m. Then the effective H_m for purposes of the Gaussian model is considered terrain following, but level above any terrain depression below stack base. (see Figure 5).

In terms of predictive errors attributable to model uncertainties in the relative values of H_a and H_m, the nocturnal period is likely to be the critical period. It is apparent that the current set of model formulations for nocturnal H_m will need to be improved. It is clear that studies of nocturnal mixed layer evolution is an active research area and any formulation and even definition is apt to be controversial. Nevertheless, efforts to incorporate state-of-the-art models for H_m through a literature review and assessment program is highly recommended. The use of accoustic sounders to obtain nocturnal mixed layer depths for direct input into model schemes may provide an attractive alternative approach.

IMPACT is a numerical grid model coded in modular form to treat separately, the transport wind field, diffusivity, plume rise, stability, and chemistry. The current version of IMPACT utilizes WEST, an objective analyses terrain dependent, three dimensional and divergence free wind field model. Input wind observations are projected upward by power law extrapolation, unless sounding data are available. Atmospheric stability data are interpolated with a r^{-2} weighting, which in turn controls the transparencies of the horizontal or vertical grids. Initial horizontal wind fields are obtained using a $1/r^2$ weighting; then the interpolated winds are vertically shifted to clear the terrain. Utilizing the gridded transparencies, the wind fields are then made divergence free. At this state, we have no clear discussion on the sensitivity and general applicability of the relationship of stability to transparences. Apparently, the table lookups for transparency values have been predetermined on the basis of prior simulations on idealized problems and are therefore only qualitatively correct for general applications. The generality of this method is thus not addressed. The discussion on the subject of transparencies is ambiguous in its presentation and its technical merit therefore could not be adequately judged. However, its sensitivity to stability is quite pronounced. Plume rise is computed by a separate module, and the current version utilizes Briggs' 1975 formulations. The module presently recognizes the first inversion layer, ignoring others, a potential source of error. Plume transport utilizes a Crowley second order flux corrected scheme which apparently is sensitive to the choice of the direction of the flux correction relative to the the actual direction of the flow field. Typically, numerical diffusion is minimal, and in general, overpredictions are prevalent. Alternative diffusivity schemes are available by selection of choice of module. The one adopted here uses a K diffusion dependent on wind speed, σ_ε, the standard deviation of the vertical wind and ℓ, a turbulent length scale, each dependent on stability in some manner. Aside from possible individual module parametric limitations and requirements for a practical code, the model user must find the appropriate compromise between desired accuracy which is a function of grid resolution and the high cost of running the model. It is clear that finer resolution exacts a very large cost increase to derive a solution. Also, the modelers report large model inaccuracies in the near source region, potentially affecting its application for sources in deep valleys and canyons adjacent to the sources. The wind field is computed elegantly and physically consistent with mass conservation, but is not valid for situations where winds are transient and nonuniform such as associated with a nocturnal jet above terrain. A power law extrapolation would fail in such a situation. (This would apply equally to those Gaussian models that utilize similar vertical extrapolation procedures). One clear advantage of IMPACT is the removal of terrain adjustment requirement since the plume dispersion is solved as part of the solution. However, once again, the accuracy of WEST will depend greatly on the quality, resolution and frequency of the input wind and stability data. IMPACT does not require an explicit specification of H_m.

Summary - Technical Evaluation of Complex Terrain Models

This set of Gaussian models exhibits wide variations in model adaptation to complex terrain. The degree of realism is of course quite limited, however, PFM and RTDM introduce Froude number scaling and the use of a critical dividing streamline height which greatly enhance their scientific bases. For this reason, both models show great promise. However, both have numerous identifiable technical and modeling deficiencies that must be addressed to become more consistent and hopefully more accurate. Technically, PFM is clearly the superior model in its handling and methodology of incorporating potential flow theory into a Gaussian framework.

The Gaussian models by necessity utilize various ad hoc engineering approaches to modeling plume rise, transport winds, mixed layer heights, stability, dispersion, boundary layer structure, terrain adjustment, partial plume penetration, limits on reflections (as mathematical artifact) as well as handling input meteorological, emissions and terrain data, all of which can potentially introduce extremely large errors in predicted concentrations. With the exception of plume rise formulations, the other models' components treatment are generally archaic at best and/or scientifically deficient. These were discussed above. The IMPACT model was the only deterministic numerical grid, non Gaussian type model in this set. An attractive feature of IMPACT is its construction in modular form, which therefore permits relatively easy upgrading to permit some maintenance of a state-of-the-art status. Greater realism and sounder physical bases are potentially possible with this type of model. This model however requires some ambiguous and rather arbitrary weighting schemes (method of transparencies based on stability classes) to complete the wind field analysis, and as such this was not clearly and convincingly argued. This model is relatively expensive to run and its accuracy in point source dispersion prediction depends in part on fine space and time scale grid resolution which must be at least as small as or smaller than the plume.

Beside quick fixes to ad hoc modeling approaches, especially for the Gaussian models, there is need for such models to be able to handle fumigation, down slope flow and flow channeling effects. Additionally, improved input data from remote sensors such as acoustic sounders and lidars for mixed layer depth, plume height and range resolved turbulence intensity, and wind and wind shear is now technically possible and commercially available. Properly deployed, these improved input data and upgraded-improved models will potentially be quite general and yield far more accurate model predictions than currently available. Some of the present models provide options for user input data such as these. Such options are highly desirable.

In the next section, a limited analyses was conducted using TRC's model evaluation data. It will be shown that model errors are as much a function of inadequate modeling methodologies as of parameterization formulations for the basic physical processes.

MODEL PERFORMANCE EVALUATION

Review of the Model Evaluation Study

Performance statistics were generated and used to evaluate eight complex terrain air quality models. This study, performed by TRC, utilized data bases from two different sites, Cinder Cone and Westvaco, each of different terrain characteristics. The results of their evaluation and tables of performance statistics are reported in the TRC report by Wackter and Londergan (1984).

Cinder Cone is a rather symmetrical and simple shaped 3-D hill; data collection was limited to non bouyant plumes in periods of neutral to stable stratification, of relatively short duration but provides high resolution receptor density. The one year of data at the Westvaco site is for a tall stack buoyant plume in rugged and complicated terrain with limited spatial resolution on observations over the complete range of atmospheric stability. TRC prepared and ran these models using a common set of input data agreed to by the modelers and subsequently generated a set of extensive model evaluation statistics against the two evaluation data bases. Their analyses plan followed the recommendations of the 1980 AMS/EPA workshop on model performance evaluation to study bias, scatter, correlation and frequency distributions (Fox, 1981). TRC's study included both paired and unpaired comparisons, on highest, second highest and one, three, and 24 hour averages. The result is a very extensive matrix of comparisons.

TRC's documentation on scope, rationale, means of performance and data interpretation was well done. There is little quarrel with their general interpretation, discussion and conclusions. One finds this analysis a useful start towards an evaluation of these complex terrain models, and moreover, one has the published results with which to conduct further comparative analysis. Overall, this was an important study which led to the result that this set of complex terrain models performance against these two data bases ranges from a qualified good to very poor. In general, TRC finds this set of models to overpredict the Westvaco data ranging from factors of 2 to 20! The analyses suggest the importance of properly modeling source characteristics and terrain configuration because the model performance exhibited large differences between the Westvaco and the Cinder Cone evaluation statistics, with respect to overall or gross errors. Each of the models varied in their relative performance depending on the nature of the test, whether it be with different data bases, different averaging periods, with stability, wind speed, highest, second highest and so on. In consideration of the technical deficiencies discussed earlier, I would recommend against a premature ranking and judgment of these models based upon the TRC study. In my opinion, at this point, one has little confidence in predicting the outcome of an extension of these models to different terrain setting for different sources, for different geographic locations.

Rather, I would like to suggest that the TRC study provide a very useful and important beginning towards identifying inadequacies in each of these models. It is instructive and highly recommended to use these data bases and model runs in conducting detailed sensitivity analyses. It is conceivable that orders of magnitude improvements are possible with simple trouble shooting exercises. The technical evaluation discussed in Part I anticipated potentially large errors; the performance study confirms this.

Other Diagnostics

I started by performing some additional analyses of the published data to highlight potential areas of critical model sensitivity (Tables 2 through 5). It was clear, however, that the interpretation and the conclusions that could be drawn were still rather limited and that further computation and analyses will be required for adequate sensitivity tests. These will be discussed in a set of recommendations to follow.

Table 2 provides a summary for the performance category of "25 Highest", the "Highest" (paired in times and paired by station), and "All Events" paired by both time and station using only hourly averages, for both Cinder Cone and Westvaco, and for runs with and without IMPACT in terms of over or under prediction (Second Highest class was not included in this summary table, but should be similar to that of the Highest class). Values for overprediction refer to the ratio of predicted to observed while conversely, values listed as under predictions refer to the ratio of observed to predicted. At a glance, a general pattern of model overprediction is apparent. Individual models vary greatly in the accuracy of prediction, but large errors are exhibited by most of the models, especially for Westvaco. On the other hand, the numbers in parentheses represent prediction falling within a factor of two; when considering the potential modeling pitfalls, such success is rather remarkable.

In the 25 highest category, only IMPACT underpredicted, and that was limited to Cinder Cone. RTDM was consistently the most successful for the three test categories. In general, almost all models were more successful with Cinder Cone than Westvaco. This should be explored further. IMPACT exhibited the largest inconsistency between the two different data sets. COMPLEX II was the most unsuccessful. These findings are consistent with those reported in the TRC report. We notice a fairly large improvement, of order 50%, in model performances using the Westvaco data when comparing results obtained between the full data set to the more limited set applicable to the IMPACT runs. In this case, the data set was limited to 20 randomly selected days out of those sets of days with the six highest observed concentration at each of the 10 monitors. This analysis is therefore biased and as a result the overprediction will be and in fact was smaller. COMPLEX I was unexpectedly more successful than COMPLEX II in just about every category throughout all the tables since σ_y was restricted to a constant value of 22.5°, independent of stability. Other performance measures for COMPLEX I and II were similar. One suspects that the large overpredictions for COMPLEX I and II arise out of the "VALLEY-like" computations where plumes are assumed to impinge onto the surface. Interestingly,

Table 2. Complex Terrain Model Performance

Overpredict: Prediction/Observed
Underpredict: Observed/Predicted

Model	25 Highest			Highest					All Events		
				Paired in Time			Paired by Station				
	Cinder Cone	Westvaco		Cinder Cone	Westvaco		Westvaco		Cinder Cone	Westvaco	
		w/IMPACT	w/o IMPACT		w/IMPACT	w/o IMPACT	w/IMPACT	w/o IMPACT		w/IMPACT	w/o IMPACT
OVERPREDICT											
COMPLEX I	(1.59)	6.37	9.23	(1.62)	3.13	7.15	5.02	5.81	2.00	(1.29)	3.39
COMPLEX II	3.78	11.47	19.62	3.62	4.05	8.98	10.13	13.83	2.00	(1.45)	3.54
COMPLEX PFM	2.51	5.03	7.70	2.31	(1.63)	(1.42)	3.01	7.47	(1.83)	-	-
4141	2.97	(1.91)	6.20	2.81	-	(1.23)	3.06	3.97	2.33	-	-
PLUME5	2.68	3.02	7.38	2.54	(1.36)	2.01	2.67	5.78	(1.67)	-	-
RTDM	(1.02)	(1.13)	(1.67)	(1.00)	-	-	(1.09)	(1.35)	(1.00)	-	-
SHORTZ	2.11	4.16	6.89	(1.77)	(1.87)	3.55	3.21	4.48	(1.17)	-	(1.98)
IMPACT	-	9.98	X	-	3.89	X	24.13	X	-	4.43	X
UNDERPREDICT											
COMPLEX I	-	-	-	-	-	-	-	-	-	-	-
COMPLEX II	-	-	-	-	-	-	-	-	-	-	-
COMPLEX PFM	-	-	-	-	-	-	-	-	-	(1.48)	(1.48)
4141	-	-	-	-	(1.52)	-	-	-	-	3.50	(1.79)
PLUME5	-	-	-	-	-	-	-	-	-	(1.99)	(1.17)
RTDM	-	-	-	(1.00)	2.03	2.17	-	-	(1.00)	3.21	2.57
SHORTZ	-	-	-	-	-	-	-	-	-	(1.13)	-
IMPACT	(1.94)	-	X	2.00	-	X	-	X	2.00	-	X

PFM does not show much better success than the other Gaussian models in general. This may be due in part to invoking the COMPLEX I and/or II schemes for flow below HC and for unstable stability class respectively.

Comparison of the Highest paired in time shows much improved results over the 25 High. RTDM still performs well, showing skill at Cinder Cone, but underpredicts at Westvaco. When paired in time IMPACT also underpredicts for Cinder Cone. When paired by station, RTDM is the most successful. All models overpredict; IMPACT was least skillful overpredicting by a factor of more than 20 for Westvaco. Further, when paired by station, the models tended to overpredict to a greater extent than for the paired in time category.

The most stringent test is the comparison of All Events, paired by time and by station. In the mean, the prediction of this category appear to be relatively more successful than for the 25 High or Highest category. For example, more than half the model's prediction were better than a factor of two on average. This good news is offset to a large extent by the practically negligible correlation between prediction and observation for Westvaco data as shown in Table 2b. This poor correlation is of course not unexpected.

Table 3 is presented which shows in detail what TRC suggested; that the degree of success of models varied with the location of the receptors. This is examined in greater detail. From their Figure 3-3 and following the suggestion in the TRC report, the receptors were classified into five distinct categories for the reasons as discussed below:

(I) Stations (1, 3, 4 & 6) are near stack (average 0.8 km downwind from the stack), on rising terrain and elevation, slightly above stack top and impacted by winds from the NW.

Table 2b. Complex Terrain Model Performance

Pearson's Correlation Coefficients

Model	WESTVACO Highest by Station w/IMPACT	 w/o IMPACT	 All Events w/IMPACT	 w/o IMPACT	CINDER CONE Highest Paired in Time	 All Events
COMPLEX I	0.66	0.79	-0.002	0.008	0.45	0.31
COMPLEX II	0.63	0.85	0.018	0.015	0.35	0.24
COMPLEX PFM	0.60	0.82	0.017	0.008	0.51	0.31
4141	0.34	0.54	0.062	0.058	0.39	0.28
PLUME 5	0.55	0.76	0.109	0.051	0.28	0.22
RTDM	0.76	0.83	0.053	0.012	0.60	0.43
SHORTZ	0.59	0.59	0.051	0.046	0.26	0.24
IMPACT	0.57	-	-0.020	-	0.41	0.29

(II) Stations (5, 7, 8 & 9) are farther downwind (average distance of 1.25 km downwind from the stack and higher than class I but impacted by N to NNW flow. Receptors in class I and II are positioned along a ridge sloping upward from the NNE to SSW and thus presents a broadside pattern to the plume.

(III) Station 2 is about the same distance from the stack as class I (about 0.9 km), is just above stack top but presents a convex face to plumes from the SSE direction.

(IV) Station 10 is the most distant receptor at 3.5 km impacted by flow from the SW with potential channeling due to orientation of the ridge at its location. This receptor is also located at about stack top.

(V) Station 11 is 1.5 km to the NW of the stack, but situated at stack base elevation.

While this is a rather arbitrary categorization, the groupings are unique and some rather interesting results are observed. Additionally, published data permitted the analysis of the standard deviations of residuals for all events and the ratio of observed to predicted variance for the 25 Highest and is presented in Table 3b.

Class I. All the models overpredicted the 25 Highest, most overpredicted for All Events. One notes that the option for gradual plume rise was not used in COMPLEX I, II and PFM and one wonders if this factor contributed to the poor prediction skill for these receptors which are relatively close to the stacks. Class I overprediction for COMPLEX I and II models is much larger than for Class II. One wonders if the Class II receptors being higher in elevation than Class I makes it subject to default concentration of zero for some of the receptors in the plume penetration region, thus reducing the average prediction values. Note that PFM is much more skillful than COMPLEX I and II in Class I which suggest the modeling approach is much improved using potential flow theory. Reference to Table 3b shows the

Table 3a. Model Comparison by Site Classes Using WESTVACO Data

A: 25 Highest Category, Unpaired Time or Station
B: All Event Paired in Time and Station

Model	Performance Group	Site Classes I	II	III	IV	V
OVERPREDICT						
COMPLEX I	A	11.49	4.65	-	(1.26)	-
	B	5.36	-	-	-	-
COMPLEX II	A	23.62	5.84	-	5.08	-
	B	5.39	-	-	-	-
COMPLEX PFM	A	6.20	2.38	(1.00)	2.27	(1.34)
	B	-	-	-	-	-
4141	A	2.18	4.66	-	-	-
	B	-	-	-	-	-
PLUME 5	A	5.29	(1.54)	-	-	-
	B	(1.25)	-	-	-	-
RTDM	A	(1.53)	-	-	-	-
	B	-	-	-	-	-
SHORTZ	A	6.73	(1.67)	(1.20)	(1.21)	-
	B	3.03	-	-	-	-
UNDERPREDICT						
COMPLEX I	A	-	-	7.40	-	(1.65)
	B	-	(1.11)	52.00	(1.43)	18.67
COMPLEX II	A	-	-	8.15	-	2.25
	B	-	(1.14)	52.00	(1.29)	18.67
COMPLEX PFM	A	-	-	(1.00)	-	-
	B	(1.07)	2.68	13.00	2.50	5.00
4141	A	-	-	9.64	4.84	3.03
	B	(1.79)	(1.36)	52.00	28.00	18.67
PLUME 5	A	-	-	3.97	(1.04)	5.33
	B	-	3.61	26.00	4.91	56.00
RTDM	A	-	(1.34)	(1.76)	(1.29)	2.28
	B	(1.88)	4.77	12.50	2.72	9.67
SHORTZ	A	-	-	-	-	(1.47)
	B	-	(1.88)	3.27	(1.24)	4.00

Table 3b. Complex Terrain Model Variability Using WESTVACO Data by Site Category

A: σ (Residuals) All Events
B: Ratio of Variance (OBS/PRED) 25 Highest

Model	Performance Statistic	Site Category I	II	III	IV	V
COMPLEX I	A	3012	998	44	128	46
	B	0.26	0.18	11.97	30.64	0.75
COMPLEX II	A	4477	1369	44	239	45
	B	0.01	0.01	34.06	0.37	1.97
COMPLEX PFM	A	1210	703	76	116	67
	B	0.03	0.05	0.17	0.11	0.19
4141	A	594	974	44	51	45
	B	45.26	0.02	44.17	42.52	5.48
PLUME 5	A	1262	622	43	69	44
	B	0.03	0.12	5.90	0.52	21.93
RTDM	A	317	246	49	68	46
		0.34	2.24	9.36	3.69	2.39
SHORTZ	A	1157	395	71	91	58
	B	0.03	0.38	4.76	1.08	2.23

standard derviations of the residuals to be the largest for Class I, and the ratio of the variances shows the predictive variances to be much greater than those for the observed (with the exception of 4141). Thus, the prediction errors are greatest for near source receptors.

Class II. Every model in this set (except RTDM) overpredicted for the 25 highest category, and with the exception of 4141, the skill increased relative to results for class I. Interestingly, all the models underpredicted in the category of All Events. A closer examination and further analysis of model performance to explain this dicotomy is recommended. The residual in Table 3b showed much improvement in the run to run errors compared to Class I. Various difference between models were noted when comparing the variance of observed vs. predicted. The behavior of 4141 seemed anomalous when comparing Class I and II performances.

Class III. With the exception of being impacted from a different direction, the receptor in Class III is similar to that of Class I, in terms of proximity and relative height of stack to receptor. Yet, there is a large difference between the model performances. In general, the skill of PFM, RTDM and SHORTZ is good; the remaining models largely underpredict the concentrations, both for the 25 highest and for the All Event category. The data, as published, do not provide sufficient bases to explain the difference between I and III. Results as published do not permit a more detailed sensitivity analyses such as possible correlation of NW flow associated with stronger winds, lower mixing heights, different stability classes, etc. versus perhaps opposite results for southerly winds. Further, the terrain at site 2 (Class III) presents a convex face to the plume while the Class I and II terrain presents more a concave to broadside face to the plume. The ratio of observed concentration between Class I and III is 4 and 2 for the 25 high and All Event category, respectively. In contrast, the ratios of the predicted concentration between Class I and Class III are much larger, greater than two orders of magnitude for the Complex I and II for example. PFM, RTPM and SHORTZ were comparatively similar, i.e., within about a factor of 5 of the observed. Table 3b lists the standard deviations of differences between observations and model predictions, for All Event category and also the ratio of variance of the observations against the variance of the model predictions according to receptor positions. The highest value for each of the four stations in Class I and II are listed. We see illustrated very small standard deviations of the residuals for Class III compared to Class I (and Class II). We also see model variances are very small relative to observed variances for Class III but this is reversed for Class I (and II) when observed variances are so very much smaller than predicted ones. (The small range for the residuals for Class III is somewhat remarkable!)

Class IV. With the obvious exception of being farther away, the topographic features of Class IV are similar to those of Class I. Yet, the model performance skill was greatly improved for Class IV. All but three of the models overpredicted for the 25 high and all underpredicted the All Event category. Thus, one tentatively concludes that the close proximity of receptors magnifies the impact of inadequate model components treatment (such as relative terrain to plume rise height, wind speed, h, σ_θ, σ_ϕ, etc.).

Class V. The character of the model performance is totally different from that for Class I! With the exception of COMPLEX PFM in the 25 high category, the models in this set underpredicted in the mean for both 25 high and All Event category. Class V is different from I by its greater distance, lower height, and possible different synoptic climatic controls on some of the model parameters. Reference to Table 3b shows large differences in prediction variances between models for both Class IV and V. It was interesting that the difference in the variance between COMPLEX I and II for Class IV was so large. RTDM exhibited the smallest and most consistent standard deviation of residual for the various receptor classes.

The Westvaco data base can thus provide interesting and useful diagnostics on the performance capabilities of these models since the receptors are sufficiently diversified in site location in different terrain features to challenge model performance consistency. This preliminary assessment provides a crude but revealing disparity in model performance skills. The Westvaco data are underutilized in terms of sensitivity analyses.

Table 4 compares model performance for Westvaco in terms of wind speed and stability using TRC published data. (The Westvaco published data did not include results of IMPACT performance in their categories) All models overpredicted in the highest 25 category with respect to wind speed. The performance in the All Event category was better, but some models underpredicted. Model skill seem to be unrelated to wind speed class as a general rule. RTDM performed quite well for the 25 high, but its performance for the All Event category was not as good. Westvaco stability classes E and F 25 High category was overpredicted by all models, and also for the All Event category with the exception of PLUME5 and RTDM. It is not clear whether this overprediction is due to the large overprediction at the close-in receptor stations 1, 3, 4 and 6, to the SSE of the stack (Class I). A sensitivity, diagnostic type analyses would provide insight. Prediction for neutral and unstable classes were mixed. Interestingly, COMPLEX I and II greatly underpredicted for these stabilities, and greatly overpredicted the stable classes. Noteworthy, 4141 exhibited similar behavior as the COMPLEX set. In contrast PLUME5, which is similar to 4141 and the COMPLEX set overpredicted the 25 high throughout the range of stability, as did PFM and SHORTZ. (One concludes that models with seemingly similar characteristics can behave so extremely differently.) The variance about the predicted means for the various stability ranges was much greater than 3 orders of magnitude for COMPLEX I, II, and 4141. Moreover, note that an extreme example of model differences was observed in the neutral stability class. Complex I and II and 4141 predicted near zero concentrations (25 High and All Events) considerably different from that observed; but PLUME5 predicted values 4 orders of magnitude higher than for the COMPLEX I, II and 4141. PFM was relatively successful compared to COMPLEX I and II in the neutral case.

Note also that the range of the observation varied by only about 50% of the observed means for any of the stability classes.

Table 4. Model Performance Using WESTVACO Data

A: 25 Highest B: All Events

w- (<2.5 m sec^{-1})
w (2.5 - 5.0 m sec^{-1})
w+ (>5.0 m sec^{-1})

Model	Statistics Class	Wind Speed w-	Wind Speed w	Wind Speed w+	Stability A-C	Stability D	Stability E	Stability F
OVERPREDICT								
COMPLEX I	A	11.53	7.76	(1.02)	-	-	10.61	10.9[illegible]
	B	2.29	5.03	(1.28)	-	-	10.65	8.1[illegible]
COMPLEX II	A	23.68	18.13	27.85	-	-	19.42	23.[illegible]8
	B	2.44	5.17	(1.49)	-	-	11.11	8.34
COMPLEX PFM	A	9.04	5.66	4.80	2.37	4.75	8.25	7.62
	B	-	-	-	-	-	(1.58)	(1.06)
4141	A	5.22	5.15	3.49	-	-	2.86	7.32
	B	-	-	-	-	-	(1.18)	(1.50)
PLUME 5	A	6.23	6.29	18.72	3.98	9.06	4.07	(1.51)
	B	-	-	(1.68)	-	(1.45)	-	
RTDM	A	(1.59)	(1.42)	-	-	(1.40)	(1.21)	(1.45)
	B	-	-	-	-	-	-	-
SHORTZ	A	5.33	7.20	4.46	3.37	5.22	5.82	7.42
	B	(1.56)	2.43	(1.39)	-	(1.92)	2.62	2.52
UNDERPREDICT								
COMPLEX I	A	-	-	-	2.61	189.63	-	-
	B	-	-	-	(1.75	∞	-	-
COMPLEX II	A	-	-	-	3.96	151.00	-	-
	B	-	-	-	13.29	∞	-	-
COMPLEX PFM	A	-	-	-	-	-	-	-
	B	(1.27)	(1.30)	5.38	(1.58)	4.61	-	-
4141	A	-	-	-	5.35	758.00	-	-
	B	2.50	(1.25)	4.37	15.50	∞	-	-
PLUME 5	A	-	-	-	-	-	-	-
	B	(1.92)	(1.26)	-	(1.7[illegible])	-	(1.56)	5.29
RTDM	A	-	-	-	[illegible]	-	-	-
	B	3.32	(1.95)	3.78	3.20	[illegible].04	2.27	(1.7[illegible])
SHORTZ	A	-	-	-	-	-	-	-
	B	-	-	-	(1.10)	-	-	-

Another interesting situation involves PFM, which for the unstable class, uses COMPLEX II to compute ground level concentration. However, for this type of stability, PFM overpredicted, COMPLEX II underpredicted in the 25 high, and while both underpredicted in the all event category, PFM skill was much better than COMPLEX II. The expected similarity was not observed. Since procedures for computing plume rise, H, was the major difference, the difference in prediction for unstable class might be due to difference in computing H and H_m. Again, diagnostic, sensitivity analysis is needed.

Table 5 is like Table 4 but for Cinder Cone data. Cinder Cone results show considerable improvement in model performance than for Westvaco. With respect to wind speed, the models show a general trend towards diminishing model prediction errors with increasing wind speed as one might suspect intuitively.

The correlation between observation and model prediction exhibited a rather unusual behavior with respect to the wind speed classes. The typical correlation for the low and high wind speed classes ranged from about 0.4 to 0.8. The correlation for the intermediate wind speed class was nearly zero for all but PLUME5, RTDM, SHORTZ. This result does not have any immediately obvious explanation.

In terms of stability, the models overpredicted the stable classes (with the exception of IMPACT and RTDM). The performance for class C and D was exceptionally well (except for IMPACT) in striking contrast to the similar comparison using Westvaco data. One can conjecture that model performmance degrades rapidly when terrain becomes more complex and pronounced, and when source emission character differs.

Table 5. Model Performance Using Cinder Cone Data

A: All Events B: Highest

Model	Statistics	Wind Speed			Stability		
	Class	w-	w	w+	C-D	E	F
OVERPREDICT							
COMPLEX I	A	2.33	(1.50)	(1.20)	-	2.80	3.25
	B	(1.93)	(1.05)	-	-	2.00	(1.95)
COMPLEX II	A	2.50	(1.60)	(1.20)	-	3.00	3.25
	B	4.34	2.36	(1.68)	(1.29)	4.73	4.91
COMPLEX PFM	A	2.33	(1.00)	-	-	2.60	2.75
	B	2.90	(1.18)	(1.00)	(1.19)	2.42	3.41
4141	A	3.00	(1.60)	(1.60)	(1.00)	3.20	4.50
	B	3.41	(1.59)	(1.41)	-	2.85	4.82
PLUME 5	A	(1.83)	(1.40)	(1.20)	(1.13)	2.40	(1.75)
	B	2.90	(1.91)	(1.55)	(1.65)	3.73	2.05
RTDM	A	(1.00)	-	-	-	(1.00)	(1.50)
	B	(1.10)	-	-	-	-	(1.09)
SHORTZ	A	(1.17)	(1.20)	(1.20)	-	(1.20)	2.00
	B	(1.48)	(1.77)	3.23	(1.10)	(1.50)	2.91
IMPACT	A	-	-	-	-	-	-
	B	-	-	-	-	-	-
UNDERPREDICT							
COMPLEX I	A	-	-	-	(1.14)	-	-
	B	-	-	(1.38)	(1.14)	-	-
COMPLEX II	A	-	-	-	(1.14)	-	-
	B	-	-	-	-	-	-
COMPLEX PFM	A	-	-	(1.20)	(1.14)	-	-
	B	-	-	-	-	-	-
4141	A	-	-	-	-	-	-
	B	-	-	-	(1.03)	-	-
PLUME 5	A	-	-	-	-	-	-
	B	-	-	-	-	-	-
RTDM	A	(1.00)	(1.25)	(1.25)	(1.33)	(1.20)	-
	B	-	(1.16)	(1.57)	(1.07)	(1.04)	-
SHORTZ	A	-	-	-	-	-	-
	B	-	-	-	(1.67)	-	-
IMPACT	A	2.00	2.50	2.50	2.67	(1.67)	2.00
	B	(1.61)	3.14	3.14	3.10	(1.73)	(1.46)

SUMMARY AND RECOMMENDATIONS

Technical Evaluation

In my opinion, the introduction and explicit use of both Froude number scaling and potential flow theory complimented with fluid modeling studies for simple, two and three dimensional obstacles provide an important extension of the physical basis for Gaussian model adaptation to terrain of varying complexity. In this respect, the PFM and to a lesser extent RTDM are potentially the most advanced and general of the Gaussian models in this set. In its current state, PFM "patches in" the computational procedures of COMPLEX I or II in certain instances and is thus a limitation to the PFM model. This limitation can and should be corrected.

RTDM doesn't incorporate potential flow parameters explicitly. However, RTDM's terrain adjustment procedures are simplistic engineering approaches that only crudely approximate the terrain adjustment predicted by a potential flow model. With the exception of terrain adjustment RTDM is relatively the most multi-featured, detailed, operational Gaussian model for complex terrain in this set, as shown in Table 1, and its relatively superior performance with Westvaco and Cinder Cone reflects this. (It was the only model, for example, that limited the amount of reflected concentrations to that imposed by the second law of thermodynamics.)

I would judge the merits of SHORTZ, 4141 and PLUME5 to be technically comparable, but deficient for their general application to complex terrain. Their terrain adjustment procedures, while clever, are without scientific merit when compared to PFM and RTDM. The relative performance of these three models would depend on the application; the source, meteorology and type of terrain and geography. Different but potentially large prediction errors would arise out a combination of unique, subtle but significant, and different model inconsistencies as discussed earlier. COMPLEX I and II have purposefully been modeled for screening applications and are therefore characteristically conservative. Thus, their technical bases for plume rise, mixing height, dispersion are not in general dissimilar from that of the SHORTZ, 4141 and PLUME5 set, but operational characteristics were definitely biased. Even here, the variation in the prediction error while large would depend on the application. For example, the performance for near source receptors at low levels will differ for those receptors at very high altitudes when plume rise and mixed large heights are low. This difference is not entirely due to the inadequacy of modeling the various parameters, but to operational model assumptions such as the artificial off-on switch for receptors at elevations about plume rise or H_m. Thus, large errors, and variances may appear in the model outcome.

Several major technical deficiencies felt to be critical were evident. Provisions for partial plume penetration into elevated stable layers were missing in all models with the exception of RTDM. The current RTDM model version, however, requires updating to reflect current knowledge about this process. Universally, the Gaussian models ignore the potentially important process of fumigation. All the models suffered in not incorporating the recent advancements on dispersion parameterization and mixed layer growth model formulations in convective conditions. Further, all formulations of nocturnal mixing heights are crude, and their current usage probably contributes to the great uncertainties in predictive skills. Those models that utilized on-site wind, turbulence, mixed layer heights are considered significant advancements in comparison to models limited to extrapolated values.

The IMPACT model shows promise. The modularity of this modeling system will permit IMPACT to attain state-of-the-art status with relative ease as the scientific and technological bases for parameterizing processes in plume dispersion improves. There is concern, however, that the current scientific bases for plume dispersion in convective conditions is not state-of-the-art. Further, the wind flow computation scheme which uses an ill defined "Method of Transparency" was not adequately justified. Additionally, model accuracy for complex terrain applications come at a heavy expense in computation costs.

This review did not examine the model performance for stagnation conditions. It is anticipated that such scenarios can contribute to larger near source impact; the knowledge of such conditions is presently insufficient to permit a technical evaluation. Diagnostic studies, however, can be developed to assess the importance of stagnation conditions on air quality, and even with reference to the EPA standards.

The skill of any dispersion model will be affected by the quality of the input meteorological data. However, it was out of the scope of this review to perform an analysis of model sensitivities to input meteorological data for this set of models. I would recommend that such analyses be performed on these candidate models for completeness. since it is anticipated that the predictive accuracy of different models will vary for a given uncertainty in input data. For example, it can be shown that spatial correlation between predicted and observed maximum ground level concentration degrades rapidly for small wind direction uncertainties. Given level and rather uncomplicated terrain, the spatial correlation for complex terrain conditions is expected to be even more sensitive to wind direction uncertainties.

Performance Evaluation

The performance evaluation conducted by TRC was extensive in the type of performance statistics generated. Their documentation of and the model prediction runs were clear and in general fair to each model. (It was unfortunate that transition plume rise was removed for COMPLEX I, II, and PFM, and that the partial plume penetration options for RTDM were excluded in the model runs). The results were mixed between the comparison for the two different data bases. Further, the results were mixed for different evaluation classes such as 25 High, Highest paired in time etc. Therefore, the conclusion drawn with the present evaluation effort was limited, but aptly discussed by TRC.

In my opinion, I do not believe it would be appropriate nor constructive to make judgment on acceptability or ranking of any of these models without additional sensitivity tests.

Using TRC published data, it became quite evident that model skill varied widely and in a generally inexplicable fashion when surveyed as a function of stability, wind speed, height, range and unique topography of receptors and for bouyant and nonbuoyant sources. Model's performances changed from overprediction to underprediction or vice versa when evaluated against different criteria classes such as 25 High, Highest, and All Event categories. Thus, the order of ranking of each model according to skill varied with the performance measure. It should be mentioned, however, that the reviewer conducted such an exercise, and did find RTDM to be the most consistent successful performer for most of the categories surveyed. However, this is not intended to imply acceptance of the RTDM approach since RTDM exhibited some erratic performance, and has some technical limitations, discussed earlier. Rather, it optimistically points out that there is clearly the possibility for the other models to improve their skills, even within the framework of the Gaussian dispersion formulation.

The wide disparity between the rankings of each model for different comparisons support this view. I believe it would be far better to extend this study to include sensitivity analyses. I suspect that even small changes in model operation procedure can drastically change the relative performance for example. This sensitivity analyses even if it were limited to the Westvaco and Cinder Cone data bases could still provide important diagnostic information or clues to "troubleshoot" each model. A number of studies that come to mind are to display the observed and predicted data in terms of season, time of day, for the Highest, the 25 High and then to stratify the All Event category, also by time of day, and by station. Parameters to be displayed should include H_m, HC, H, and stability class as well as the observed and predicted concentrations. Separate scatter plots and analysis of observed, predicted and differences of observed and predicted, and ratio of overpredicted or underprediction as dependent variables against stability, wind speed, H_m, HC, H, H_m-H, H-HC, time of day, and distance are possible diagnostic tools. Perusing Tables 2-5 and examining data on correlation, positive residual etc. showed such erratic model performance and raised too many unanswered questions which cannot be resolved with the currently available and published tables. These studies should attempt to separate out errors due to model operations versus those due to technical or scientific limitations. It is strongly recommended that modelers contemplate the introduction of Froude number scaling and potential flow adjustments for complex terrain applications.

Future Prospects and Recommendations

It is recommended that:

1. Models be upgraded to include those aspects of potential flow theory and Froude number scaling such as described in the PFM model.

2. Improve the dispersion parameterization system by a) upgrading with state-of-the-art convective scaling methodology and b) use on-site turbulence data. For example, the use of remote sensing systems such as doppler sodars may be promising. In this regard, the turbulence intensities will be more representation of the dispersion potential at plume rise.

3. Improve the transport wind data with the requirement for on-site mean vector wind profiles. The local transport wind direction and shear may be totally unrelated to any extrapolations from alternate and distant measurements, such as airport observations for example.

4. Improve the methodology for determining the mixed layer depths, especially during the nocturnal periods.

5. Include provisions for partial plume penetration into elevated stable layers using the theory of Briggs (1984) or Weil and Brower (1983).

Tools to develop improved, more general, theoretically sound and accurate Gaussian models are currently available. These include remote sensing radars, sodars and lidars to provide high resolution, accurate, unambiguous data on

a) wind speed, wind direction, shears
b) mixed layer heights
c) plume characteristics
d) ambient turbulent wind fluctuation statistics, and turbulent intensities
e) multiple layers

Carefully utilized, these sensors can yield information leading to better terrain adjustment procedures for example. Further, differential absorption lidars are available for some pollutants from which range resolved concentration data can be obtained for research and model validation purposes.

Obviously, if very high concentrations can occur due to fumigation processes, some research and model development effort will be necessary to incorporate this process in the computation schemes. Quick fixes may be possible to the Gaussian model through a source enhancement technique (analogous to a source depletion concept), but the truly improved model will require substantial effort.

Active field and fluid modeling programs are underway (ASCOT; the EPA Complex Terrain Study, the EPRI Plume Model Validation and Development Studies, and other less comprehensive ones) to address the issues on Complex Terrain modeling such as plume impaction and terrain adjustment, flow channeling, dispersion, and data resolution requirements. Advancement in complex terrain modeling will certainly benefit by a strong effort to follow through those basic studies to the point of model development and validation.

REFERENCES

Benkley, C.W., and L.L. Schulman. Estimating hourly mixing depths from historical meteorological data. J. Appl. Met. 18: 772-780, 1979.

Briggs, G.A. Plume rise predictions. In: Lectures on air pollution and environmental impact analyses. AMS, Boston, MA, pp 59-111, 1975.

Briggs, G.A. Plume rise and buoyancy effects. In: Chapter 8 in Atmospheric Science and Power Production, DE 84005177 (DOE/TIC-27601), NTIS, U.S. Dept. of Commerce, Springfield, VA 22161, pp 327-361, 1984.

Fox, Douglas G. Judging air quality and performance. Bulletin of the Am. Meteor. Society. 62(5): 599-609, 1981.

Frank, N.H., B.E. Blagan, and H. Slater. Diurnal patterns of sulfur dioxide. Presentation at 74th Annual Meeting, APCA, Philadelphia, PA, 1981.

Gudliksen, P.H., and M.H. Dickerson. Executive Summary: Atmospheric Studies in Complex Terrain, Technical Progress Report, FY 79-83. UCID-18878-83. Summary, ASCOT 84-2, 1983.

Hilst, G.R. Plume Model Validation. EPRI EA-917-59, Electric Power Research Institute, Palo Alto, CA, 1978.

Holzworth, G.C. Mixing heights, wind speeds and potential for urban air plollution throughout the contiguous United States. AP-101, EPA. Research Triangle Park, NC. 118 pp, 1972.

Hunt, J.C.R., and R.J. Mulhearn. Turbulence dispersion from sources near two-dimensional obstacles. J. Fluid. Mech. 61: 245-274, 1973.

Hunt, J.C.R., Puttock, J.S., and W.H. Snyder. Turbulent diffusion from a point source in stratified and neutral flows around a three-dimensional hill. Part I - Diffusion Equation analyses. Atm. Environ. 13: 1227-1239, 1979.

Hunt, J.C.R., and W.H. Snyder. Experiments on stably and neutrally stratified flow over a model three-dimensional hill. J. Fluid Mech. 96: 671-704, 1980.

Lamb, R.G. A numerical simulation of dispersion from an elevated point source in the convective planetary boundary layer, Atm. Environ. 12: 1297-1304, 1978.

Lamb, R.G. The effect of release height on natural dispersion in the convective planetary boundary layer. In: Proceedings of the Fourth Symposium on Turbulence, Diffusion and Air Pollution, Reno, NV., AMS, Boston, MA, 1979, pp 27-33.

Lamb, R.G. Diffusion in the convective boundary layer. In: Atmospheric Turbulence and air pollution modeling, F.T.M. Nieuwstadt and H. Van Dop, eds. D. Reidel Publishing Co., Dordrecht, Holland, 1984, pp 159-230.

Lavery, T.F., A. Bass, D.G. Strimaitis. A. Venkatram, B.R. Greene, P.J. Drivas, and B.A. Egan. EPA Complex Terrain model development program. First milestone Report, 1981. EPA-600/3-82-036, U.S. Environmental Protection Agency, Research Triangle Park, NC, 1982, 305 pp.

Lavery, T.F., D.G. Strimaitis, A. Venkatram, B.R. Greene, D.C. DiCristofaro, and B.A. Egan. EPA Complex Terrain Model Development, Third Milestone Report - 1983. U.S Environmental Protection Agency, Research Triangle Park, NC. 1983, 271 pp.

Report to the U.S. EPA of the Specialist Conference on the EPA Modeling Guidelines, Feb. 22-24, 1977, Chicago, Ill.

Smith, M.E. Review of the Attributes and Performance of 10 Rural Diffusion Models. Bulletin of the American Meteorological Society 65(6): 554-558, 1984.

Strimaitis, D.G., A. Venkatram, B.R. Greene, S. Hanna, S. Heisler, T.F. Lavery, A. Bass, and B.A. Egan. EPA Complex Terrain Model Development, Second Milestone Report - 1982. EPA-600/3-83-015, U.S. Environmental Protection Agency, Research Triangle Park, NC. 1983, 375 pp.

Strimaitis, D.G., T.F. Lavery, A. Venkatram, D.C. DiCristofaro, B.R. Green and B.A. Egan. EPA Complex Terrain Model Development, Fourth Milestone Report - 1984. U.S. Environmental Protection Agency, Research Triangle Park, NC, 1984, 319 pp.

Uthe, E.E. Cooling tower plume rise analyses by airborne lidar. Atm. Environ., 18(1): 107-119, 1984.

Wackter, D.J., and R.J. Londergan. Draft: Evaluation of Complex Terrain Air Quality Models. TRC Project 2465-R81, TRC Env. Cons. Inc., E. Hartford, CT. Contract No. 68-02-3514. U.S. Environmental Protection Agency. Research Triangle Park, NC 27711.

Weil, J.C. Applicability of stability classification schemes and associated parameters to dispersion of tall stack plumes in Maryland. Atm. Environ. 13: 819-831, 1979.

Weil, J.C., and R.P. Brower. The Maryland PPSP dispersion model for tall stacks. Prepared by Environmental Center, Martin Marietta Corporation for Maryland Department of Natural Resources (Ref. No. PPSP-MP-36), 1982.

Weil, J.C. and R.P. Brower. Estimating convective boundary layer parameters for diffusion applications. Prepared by Env. Center, Martin Marietta Corp. for Md. Dept. of Natural Resources. (Ref. No. PPSP-MP-48), 1983.

Willis, G.E., and J.W. Deardorff. A laboratory study of dispersion from an elevated source within a modeled convective boundary layer. Atm. Environ., 12: 1305-1311, 1978.

Willis, G.E., and J.W. Deardorff. On plume rise within a convective boundary layer. Atm. Environ., 12: 2435-2447, 1983.

REVIEW OF COMPLEX TERRAIN MODEL PERFORMANCE

Prepared for the AMS-EPA Steering Committee

by

Robin L. Dennis*
Meteorology and Assessment Division
Atmospheric Sciences Research Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

March 1985

---

*On assignment from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce.

INTRODUCTION

This review is divided into three main sections: Perspective on the Models; Model Intercomparison; and Review of the Evaluation Framework. The Perspective on the Models section presents a brief discussion of general knowledge as it relates to the eight models included in the performance evaluation. The intent is to provide some indication of the "currentness" of the assumptions used in the eight models. The Model Intercomparison section examines how the models seem to compare; can the differences be explained? An indicative evaluation is presented that examines the quality of the answers from the models. Suggestions for a needed improvement in the measures used in an intercomparison are discussed. Preliminary judgments are presented with respect to the performance of the models and the relation of performance to currentness of the assumptions used in the models. As well, several apparent problems with the models or with how they were run are noted. The Review of the Evaluation Framework examines the design of the TRC evaluation (Wackter and Londergan, 1984) with respect to stated and unstated goals for such evaluations. The evaluation measures used are critiqued and suggestions for additional, needed measures are given. Issues involved with transferability of the results to other complex terrain problem domains are discussed.

PERSPECTIVE ON THE MODELS

Interestingly, the parameterizations embodied in the eight complex terrain models cover a fairly broad range of approximations of the physical process involved in point source dispersion. There is a range in the sophistication of the approximations to flow over or around a terrain feature. There is a range in methods to estimate turbulence parameter magnitudes and in the number of modifications to them that are included. The use of local turbulence data as the basis for estimation is an important facet. Most of the models seem to be blithely run in violation of the second law of thermodynamics. Plume-rise algorithms are similar across the models, but implemented differently for the short downwind distances involved.

Complex Terrain Flow

The complex terrain features that are considered in this evaluation are (1) a sloping ridge and (2) a hill. These features are two generic and "simple" types of complex terrain. The pollutant source is a single elevated point source, fairly close to the terrain feature. For complex terrain features, such as these, it is known that the mean path of the plume of pollutants does not follow the mean streamline (Hunt and Mulhearn, 1973; Egan, 1975). Two models, however, IMPACT and SHORTZ ignore this element of the physics.

One has the distinct impression from reading the documentation that the developers of IMPACT do not believe that pollutant plumes necessarily follow the mean streamline. It appears as if the influence of the terrain features is to simply constrain the wind not to blow through the terrain surface in the calculation of a non-divergent wind field. The entire link to inhomogeneous flow and the resulting effect on plume displacement and on lateral and vertical diffusion rates caused by obstacles seems to be lacking. Evidentally a wind model, MATHEW, closely related to the one used in IMPACT (WEST) has extremely unrealistic and undesirable behavior at the leading edge and at the top of terrain features where there is compression. This behavior results because continuity of mass flow along the streamline is not checked. Making the wind field divergence-free does not take care of the problem (Graeme Lorimer, Australia, personal communication). One suspects the same problems could show up in WEST. There also seems to be no consideration of the well-known problems that come into play when using a multiplebox, K-theory model to model point sources (e.g. Lamb and Durran, 1978; Deardorff, 1978).

SHORTZ also seems to ignore the impact of an obstacle on streamlines and on the mean path of the plume. Plume impaction is not allowed if the turbulence intensity is less than 0.01, but such low intensities rarely occur in these data. Thus, the plume is essentially always assumed to be in the mixed layer. The documentation states that when in the mixed layer the plume height remains constant with respect to sea level, once it has reached final rise, regardless of terrain height. If terrain height is higher than the stabilized plume height, then plume height is fixed at zero. Thus, lift of the plume over an obstacle seems not to be considered. Citing the second law of thermodynamics and avoidance of "unrealistic compression of the plume," the effective mixing depth is terrain-following for obstacles. This latter treatment of mixing depth does seem consistent with the more recent observation by Hunt, Puttock and Snyder (1979) that, qualitatively, the variation of concentration on a 3-dimensional hill is primarily determined by displacement of the streamline, rather than by convergence or divergence of the streamlines. It is not apparent the model developers had this in mind, however.

The simplest approximation to the physics underlying the above observation that plumes do not follow the mean streamline is more than ten years old (Briggs, personal communication). In neutral and stable conditions, the plume loses part of its effective stack height relative to the surface of the terrain feature. This simplest approximation is to assume that the loss in height of the plume is half the full height that would have been calculated had not the feature been present (half-height correction). In stable conditions, the approximation is that the plume will maintain a constant elevation, irrespective of the terrain feature. If the terrain feature is high enough, the plume will impact. Three mode[illegible]OMPLEX I, COMPLEX II, and PLUME5, treat the flow over or around the terrain feature in this basic manner.

It has been recognized during the last 10 years that this simplest approximation is terrain-feature dependent (ANL, 1977). The half-height correction is most appropriate for terrain objects with roughly equal horizontal and vertical dimensions. For two-dimensional ridges, due to distortion effects, it appeared that a half-height correction would be too conservative. It was suggested that a terrain-following trajectory might be more appropriate (Egan, 1975).

The next level of sophistication is use of the dividing-streamline concept to determine whether a plume will rise over the top of or impact the surface of the obstacle in neutral and stable conditions (Hunt, Puttock and Snyder, 1979; Snyder, Britter and Hunt, 1980; Snyder, 1983). The use of the Froude number to define a critical height is an important advance in approximating the physics of plume behavior in complex terrain, even though it is representative of simplified cases. Two models, RTDM and COMPLEX/PFM, use the critical height concept. The application is more "exact" for neutral stability (based on potential flow trajectory solutions) and more empirically based for stable conditions (based on the results of experiments at the EPA Fluid Modeling Facility).

Giving the model developers the benefit of the doubt, it appears that one model, 4141, attempts to "simulate" the physics of the dividing-streamline concept by using a 1/4-height correction to the plume height in stable conditions, rather than determining whether or not the plume will go over the hill or impact it. The documentation, however, does not imply that the 4141 model developers had such a "sophisticated" rationale in mind.

A further increment of sophistication is to actually account for, in a single model, the observation that the shape of the terrain feature is an important determinant of the plume height above the surface when the plume has enough kinetic energy to surmount the obstacle (Hunt, Britter and Puttock, 1979). That is, the half-height correction, even used with the Froude number, is too simplistic; it does not reflect the sensitivity of plume height to terrain geometry and meteorological variations. COMPLEX/PFM incorporates a first-generation, first-order approximation to account for this sensitivity. Thus, the PFM component of COMPLEX/PFM represents the most complete operationalization of current understanding of plume behavior in interaction with "simple" terrain features for "straight-forward" meteorology. This would appear to be no small achievement.

There is still a lot of "physics" that the models included in this evaluation are not able to consider, such as being able to account for the fact that plumes can behave differently for the same Froude number or the fact that upwind boundary conditions can greatly affect the gross flow features of the wind field upon which the local terrain influence is superimposed, possibly dominating or suppressing local terrain influences. The evaluation sites seem to be representative of the more simple cases of complex terrain and wind field flows, however. This is clear for Cinder Cone Butte, but less clear for the case of Westvaco. Thus, there does not seem to be an inordinate disparity between the simplified real world against which the models are being compared and the simplified model assumptions. Even though more complex situations, which are closer to many of the real world applications, are not being examined and tested in this evaluation, the range of physics represented by the assumptions in the different models should be very informative, both from a scientific and a regulatory point of view.

Dispersion Parameters

Essentially three sets of dispersion estimates are used in the models. The majority use the Pasquill-Gifford-Turner (PGT) curves (COMPLEX I, COMPLEX II, COMPLEX/PFM, 4141, PLUMES). However, in 4141 the lateral dispersion coefficients are increased by a factor of 1.82 over the PGT values, citing "considerations of sampling time." Apparently this was done for all classes of stability. For two models, RTDM and SHORTZ, on-site turbulence data are used to estimate the turbulence parameters. Inclusion of on-site turbulence data is expected to be a real advance. IMPACT uses an empirical set of estimates, which assigns lateral turbulence values to exogenously determined classes of stability in order to calculate both horizontal and vertical dispersion parameter values. This approach seems backwards, but similar in spirit to the use of PGT curves.

An investigation of the differences between the use of the PGT curves versus the local turbulence data, using a small non-random sample of Cinder Cone data (20% of the data), indicates the following: (1) There is essentially no difference between the estimates of vertical dispersion for stable conditions at a one-kilometer distance using either PGT curves or on-site turbulence data (and RTDM's algorithm). These estimates agree well with values presented in recent literature (Irwin, 1983). (2) There is a minor difference (15-20%) between the estimates of horizontal dispersion of Cramer (SHORTZ) and RTDM at one kilometer for neutral and stable conditions and these estimates are in good agreement with the most recent literature (Irwin, 1983). However, the estimates of horizontal dispersion, based on local turbulence data, are a factor of 2 to 4 larger than the estimates based on PGT curves. The difference is roughly a factor of 2 for neutral conditions. The difference is roughly a factor of 3-4 for very stable conditions. Thus the differences in the estimation of the lateral dispersion parameters should have a large influence on the relative predictions of the models.

All of the Gaussian-based models include plume buoyancy corrections to account for the fact that the dispersion is enhanced for strongly buoyant plumes. Thus, all of the models take into account what seems to have been an earlier concern that models were not including corrections for buoyancy (ANL, 1977). Somewhat different approaches are used in SHORTZ and 4141 from the "standard" approach used for the other models based on guidance from Pasquill (1976). The difference seems minor in the case of 4141.

Three of the models, RTDM, PLUME5, and SHORTZ, include enhancements of the lateral diffusion due to wind shear. The suggested form from Pasquill (1976) was followed.

Second Law of Thermodynamics

If a flow conserves mass, then concentrations in that flow cannot increase with increasing travel time (a simple statement of the second law of thermodynamics). Using the conventional reflection algorithm of the older Gaussian models doubles the plume center-line concentrations for cases of plume impaction on a terrain surface (a violation of the second law). In essence, the model calculation, because it is steady state, assumes that ground-level has been at the height of the plume center-line the full distance from the source. This assumption of full reflection of the plume is rather unrealistic for complex terrain models, a point that has been made both mathematically and empirically (Hunt, Puttock and Snyder, 1979; Hunt, Britter and Puttock, 1979; Snyder, Britter and Hunt, 1980). The difference between what seems to be a more reasonable concentration estimate and an estimate based on full reflection is roughly a factor of two.

Only one model, RTDM, did not apply the full reflection assumption. A method is employed in the RTDM that produces a simplified, conservative estimate of the center-line concentration at the time of impaction. That is, the "real world" answer is expected to be lower than their estimate due to their not considering dispersion that still can occur and their not considering both wind meander effects and the build-up of eddies that keep the plume from fully impacting the surface.

One would, therefore, expect the other six Gaussian models to systematically overpredict concentrations associated with impaction. They should also systemically predict higher concentrations than RTDM for the extremes. While it is nice to establish this "known" fact for regulatory purposes, it is not particularly illuminating when one is trying to understand how well the models could predict.

Three models, COMPLEX I, COMPLEX II, and COMPLEX/PFM, could have been run in a mode that established limits on the center-line concentration, similar to those used in RTDM, by setting user option IOPT(25)=5. They were run in the regulatory mode, however, employing full reflection, setting IOPT(25)=1, due to regulatory needs. There is no question that it is very useful to examine the performance of these models as they are used in a regulatory, screening mode. One obtains a clearer understanding of the margin of safety that is built into the predictions produced by these models for the screening mode. However, to evaluate the science in the models, especially COMPLEX/PFM, the models should also have been run with

IOPT(25)=5, because a simple "correction" of the results is not possible without considerably more information than that included in the evaluation. It is unfortunate that these models were not run both ways.

Plume Rise

The models use plume-rise formulations that are from or consistent with the 1975 set of Briggs' formulas (Briggs, 1984). There is a noteworthy aspect in this evaluation associated with the plume-rise calculations. While the models all cluster around a few formulas of plume-rise, a major difference is whether the models use final rise or a gradual rise. A back-of-the-envelope calculation implies that the difference could be important.

The Westvaco stack height is 183 meters. The monitors vary in distance from the stack, ranging from 800 meters to 1500 meters from the stack along the ridge. These distances are all less than 10 stack-heights from the source. At 1 kilometer the difference between a final rise and a gradual rise plume height is 50% for neutral stability, assuming final rise is achieved at a distance of 1.8 km. This effect should not affect the computation of maximum concentrations, because one would assume that plume impaction is the cause of the maximum. It will affect predictions of concentrations which are lower on the concentration distribution, however, and would be expected to produce a bias toward underprediction, especially for the lowest concentrations. This bias will also be monitor (distance) specific for those models that use final-rise plume height. Several models had options for gradual rise, but were run using final rise. It is not clear from the TRC documentation why this choice was made.

MODEL EVALUATION AND INTERCOMPARISON

One problem that perenially faces model evaluations is that the data sets on which the evaluations are based are always too limited and there are never enough data sets to answer all of the relevant questions. Therefore, the evaluations must be used to the hilt. In this case, diagnostic evaluation becomes an important part of any overall evaluation, because reasoned inference will be required to assess how the models could be expected to perform and compare in situations other than those represented by the specific evaluation site data set being used. One would assume that an evaluation oriented towards regulatory application of the models would want to know whether or not a good or bad showing of a particular model on the evaluation data set is a fluke and whether such a showing is expected to carry over to other cases.

To provide a basis for reasoned inference, a regulatory evaluation must go beyond the simple comparison of the predicted and observed maxima. This is because, from a simple comparison one does not really know what one has, except the 25 largest numbers, for example, from several black boxes. There is no way of telling if these 25 numbers mean the same thing. One does not know how to evaluate the comparisons without going further into some diagnostic work on the numbers.

Importantly, the diagnostic work relates not only to the results from the black boxes, but also to the information contained in the data set of observed values. That is, the types of measures used, how they are specifically defined and used, and the subdivisions of the data (breakdowns) used must be developed in interaction with the type of results the models are producing and the type of behavior evidenced by the real system. Reasonably specific guidance can be established on how to begin an evaluation for particular types of pollution/meteorological systems, given their idiosyncracies. However, it is just that, only a beginning. Conducting an evaluation is an iterative process. The second and succeeding steps in the evaluation depend on the first. The first step can be the most mechanical of all, but it is only the first step. What is presented in the TRC document, therefore, is only the (partial) first step towards an evaluation. An actual evaluation is far from being realized.

Examination of Distributions/Populations

To begin with, it is clear that in many instances different, possibly non-comparable populations are being compared, especially for comparisons based on the Westvaco data set. For one model, the population of the top 25 apparently corresponds to F stability and station number 1; for another model the top 25 seems to represent a population from D stability and station 6; for another model the top 25 represents a population from F stability and stations 9 and 5; and for yet another model, the top 25 represent a population from D and F stabilities over several stations. The question is, do these populations have anything in common, are they even comparable?

The second question is, what do these populations of model predictions have in common with "reality?" For the observed data, the top 25 represent a population from D and F stabilities over several stations (most likely 4). Thus for "reality" there are 16 subpopulations of stability and station. To obtain the 25 highest, one is, at most, picking only the top two from each subpopulation. This population of extremes, made up of several subpopulations, is then being compared with a population made up of the top 25 from a single subpopulation.

There is no reason to expect the population of the predicted "top 25" to be the same as the observed "top 25" nor for their distribution of concentrations to be the same. One only has to remember that a distribution of concentrations which is the sum of two log-normal distributions is not itself log-normal. However, the TRC evaluation report assumes that all of the "top 25" populations are similar across models, and can validly be compared, rather than establishing that fact for the data set in hand. In fact, one might expect RTDM to produce a flatter cumulative distribution of this "top 25" compared to the cumulative distributions produced by the other models. That is, the range between the highs and the lows of the top 25 could be less for RTDM, because several combinations of meteorological conditions (subpopulations) contribute to its top 25 predictions. Thus, for RTDM, a flatter distribution than those produced by the other models could be a "correct" answer, instead of expecting the models to produce the same distribution. One can try to crudely estimate the relative spread of the distributions coming from the different models for Westvaco using the eight stations on the ridge (one can't do it properly, given the information available in the TRC report) and then for Cinder Cone (again in a very crude manner). To do this, a ratio of average concentration using a small number of observations to an average using a large number was computed. The following comparisons result:

Table 1. Ratio of Maximum Values for Different n

Westvaco: n=8/n=200 (mean of station maxima/mean of station top 25, across 8 stations)

Cinder Cone: n=25/n=104

	Westvaco	Cinder Cone
COMPLEX I	1.50	2.5
COMPLEX II	1.92	2.7
4141	2.59	2.7
RTDM	2.34	2.7
PLUME5	3.46	2.7
COMPLEX/PFM	3.48	2.8
SHORTZ	2.16	3.0
IMPACT	---	2.6
Observed	1.88	2.5

As expected, there are a variety of ratios for Westvaco. It has still to be established how comparable one should expect the slopes (ratios) to be. It is noteworthy that the RTDM slope is not the lowest (i.e. not the flattest cumulative distribution) and it is also higher than the slope for the population of observed values. (Whether it is significantly different should be established.) One would have expected the observed and RTDM populations to be the most similar of all, if RTDM is predicting well. For the Cinder Cone results, the ratios are much more comparable across the models. It is quite possible that the populations represented in the Cinder Cone data are fairly comparable in a relative sense. If so, then all the models, except SHORTZ, do a very good job of representing the relative spread of the distribution of high concentrations in the Cinder Cone data. The degree of agreement on the Cinder Cone data is rather phenomenal. One might suspect that differences across models, as evidenced by the differences in bias, are due to systematic, multiplicative factors associated with parameterization of physical process. (A more adequate treatment of Westvaco data appears later in the review due to "finding" some data late in the review process.)

For the Westvaco data, one could have a model that underpredicts drastically at every station, but one, and for every stability, but one. For that one station and one stability category, the bias of the "top 25" is only a factor of two greater than the observed "top 25." Is that model as good as RTDM because it has the same bias? Given the present level of information contained in this evaluation, one has no "objective" basis for making a judgment in answer of that question that could ever extend beyond the very specific, local population of the "top 25 from all stabilities and all stations 800 and 1600 meters from the stack on a ridge exactly as Westvaco's for the year's meteorology represented by the given data set."

Thus, to compare the model predictions, with the "real world" and across models, one should use populations that are as similar as possible. This means that different breakdowns of the data must be developed, based in part on how the models tend to predict. Exploration and display of the different sub-populations could be achieved by the use of histograms and box plots. The box plots have the advantage of containing more information than standard deviations; box plots can present the median, the skewness, the width of the distribution and odd-ball values, all displayed at once. Only with such information can relevant subsets of data be constructed that will produce well-defined and interpretable comparisons. Of course, this assumes that the actual data are available in order to generate all of these possible plots, tables and diagrams.

Once this diagnostic work has been done, one can then go back and contruct a careful and meaningful set of results that meets the needs of a regulatory evaluation. The TRC report represents a good first iteration of the many iterations required to develop a good evaluation of these models.

Following a different vein, one should have noticed that the above results present a totally different evaluation of the ability of the models to reproduce the relative spread of the concentration distribution from that contained in the TRC evaluation report. The model comparisons look terrific for Cinder Cone and not all that bad for Westvaco (see also the section on Exploratory Analysis of Top 25). There is no possible manner in which the same conclusions can be developed from the values presented in the TRC document. That is because the measures TRC uses to compare standard deviations of residuals and frequency distributions are heavily influenced by the amount of bias in the model predictions.

For these models, the influence of the bias is so great as to make the interpretation of these measures, as used by TRC, totally misleading. In the TRC report the comparisons of standard deviations of the residuals and the frequency distributions are essentially meaningless for these particular model comparisons. To have any meaning, the computation of the measures used by TRC must include a correction for bias. One suggestion is the use of a statistical test that can account for bias, such as the

Siegel-Tukey test (Gibbons, 1976). But the shape is best measured by parameters of slope and spread. Why not "measure" these directly, rather than use the indirect manner of the TRC report? Thus, better measures could be included in the performance evaluation. In addition, one of the best ways to display this information is through the use of graphs. This evaluation is remiss in how the information is displayed.

Breakdown by Stability

One of the more useful breakdowns of the data is by stability class in order to try and understand how the models are performing. As an example of this, Table 2 presents the average of the highest concentrations within each stability class. Unfortunately, the information of this table is not directly available; one has to calculate the values from bias and observed values.

Table 2. Means of Highest Concentrations Associated With Each Stability Class

Westvaco
(micrograms/m**3)

Model	A-C n=25	D n=25	E n=25	F n=25
COMPLEX I	418	8	11,859	18,169
COMPLEX II	276	10	21,767	38,638
4141	204	2	3,200	12,205
RTDM	1,048	2,121	1,353	2,415
PLUME5	4,346	13,745	4,555	2,520
COMPLEX/PFM	2,547	7,205	9,219	12,790
SHORTZ	3,678	7,912	6,513	12,363
Observed	1,092	1,517	1,118	1,667

Cinder Cone
(micrograms/m**3)

Model	C-D n=30	E n=38	F n=36
COMPLEX I	27	51	43
COMPLEX II	40	123	108
4141	30	74	106
RTDM	29	24	24
PLUME5	51	97	45
COMPLEX/PFM	37	63	75
SHORTZ	34	39	64
IMPACT	10	15	15
Observed	31	26	22

It is unfortunate that for Cinder Cone TRC combined C and D stability, because some of the models use very different algorithms to develop their predictions for each of these stabilities. I suspect that TRC did this in order to obtain a larger sample size, n. But I would rather have been given more information about the individual samples and had them disaggregated fully in a manner that makes sense in terms of the parameterization of physical processes that influence the predictions made by the various models. If the sample size, n, is too small, so be it. Relevancy is more important in this instance than rigidly following some statistical rule. One would also like to have more information about each of the distributions to be able to adjust the means for their different sample sizes for comparisons across stability classes.

This reviewer attempted to understand the differences between the models and the differences in their predictions, listed in Table 2. This was done in the spirit of the goal of the TRC evaluation: "... a systematic evaluation of these models to decide in an objective manner which models should be included in the guidelines and what recommendations should be made concerning the use of these dispersion models for regulatory application." The second goal was also kept in mind: " The principal objective of this project is to produce performance statistics so that EPA and a group of reveivers may judge the relative merits of different models."

Several conclusions beside the obvious ones about RTDM's performance relative to the other models surfaced. One is that there is not sufficient information collected and presented as part of the evaluation data base to carry out an adequate evaluation of the models. What is done is too simplistic. Too many factors are influencing each set of numbers and these influences cannot be uncovered with the presently available information. Yet the diagnosis of what is "going on" in the numbers is important to the evaluation of the models' performance.

Another conclusion is that components of the models not directly related to their handling of complex terrain seem to have as much or even greater influence on the predictions of the models as do the assumptions about how to treat complex terrain. This is a judgment that must be based on insufficient information and relies heavily on noting consistent patterns of behavior between models. The result is, that it is difficult to establish the value of the different levels of sophistication with which complex terrain is treated in the different models. The three most obvious components that strongly influence the predictions of the models, in addition to the component reflecting assumptions about complex terrain, seem to be treatments of plume-rise, buoyancy induced dispersion and eddy diffusion.

The differences between the models on the Westvaco data for D stability, shown in Table 2, could be an indication of the importance of plume rise on the model predictions, because many use the same half-height correction and only SHORTZ should allow plume impaction. It is impossible, given the information available, to more precisely know what effect plume rise is having other than to infer that it is an important effect. For example, predictions of COMPLEX II and COMPLEX/PFM are similar in the Cinder Cone comparison, but vastly different for the Westvaco comparison. For the Westvaco data, SHORTZ may predict higher concentrations than RTDM for neutral stability, even though both models have similar estimates of lateral diffusion, because plume height above sea level is held constant in SHORTZ, whereas RTDM uses a half-height correction. Yet, why is this same pattern repeated for the Cinder Cone data set where no plume rise is involved?

One would want to be able to establish, more quantitatively, what relative influence is coming from what component of the model. One suggestion would be to include as part of the evaluation data set the center-line prediction at the distance of the receptor for each hour. This would help to develop some diagnosis of the model's predictions and help assess the importance of plume rise, buoyancy induced dispersion and eddy diffusion on the actual prediction at the receptor. This dissection of a predicted concentration is important for an assessment of the question, "is the better science in the model having a positive effect on the predictions?" (Post Script: an evaluation in this spirit was carried out by Dennis and Irwin, 1985 which provided extremely important illumination of model behavior.)

The apparent differences in lateral dispersion parameters based on on-site turbulence data compared to the PGT curves seem to be important. The differences appear to be greater for the Westvaco data set than for Cinder Cone. Thus differences seem to be a function of data set. Time and data availability did not permit the quantitative establishment of this inference.

Much of the difference between COMPLEX II and RTDM for stable conditions can be explained by expected differences in the horizontal diffusion coefficients and the reflection of the plume when the plume is expected to impact (a factor of 6-8 compared to an observed factor of 10). Differences in lateral dispersion parameters could explain a major portion of the difference between COMPLEX/PFM and RTDM for the neutral case for both Westvaco and Cinder Cone. Without knowing more precisely the difference in lateral turbulence values (PGT versus on-site), one cannot truly evaluate COMPLEX/PFM's predictions, which account for terrain shape, against RTDM's predictions, which do not account for terrain shape. The turbulence estimates used by each model at the distance of the receptor should be output and made available as part of the evaluation data set.

Differences between COMPLEX I and COMPLEX II for impaction are consistent with the difference expected as a result of the 22.5 degree sector averaging compared to a bivariate Gaussian point estimate at a one-kilometer distance. It is interesting to note that the difference is expected to increase to a factor of five at a distance of five kilometers.

It is not clear why SHORTZ and COMPLEX/PFM give such similar predictions, both for the breakdowns by stability and by monitoring station (Westvaco). Having a grasp of the reasons for their similarity on these two data sets, which are rather special, would seem to be very important to a judgment about using either model under different circumstances for regulatory purposes.

The "top 25" observed concentrations seem to be very uniform in space and across stabilities for Westvaco and have a rather simple trend across stabilities for Cinder Cone. None of the models can reproduce this space and stability class behavior of the "real world," although RTDM comes by far the closest of any model. The question that should be raised is, do the models have to be able to reproduce the spatial and stability class behavior of the real world in order to come up with useful predictions for regulatory purposes? This is where some insight into the sample "populations" for the different space and stability classes is needed, as discussed in an earlier section.

Spatial Behavior

The evaluation does not really address the spatial and stability class aspects of model behavior, even though TRC thinks it evaluates the spatial behavior by using the Pearson correlation coefficient. The reason that the Pearson correlation coefficients (Spearman, too) are not giving much information at all (Westvaco data), is because they are anchored on the lower end by stations numbers 2, 10, and 11. These stations have very low values, more than a factor of ten lower for predicted values and a factor of four lower for observed values.

This very large gap between the two clusters of points will, of course, produce a fairly large coefficient of correlation, regardless of whether or not the values are correlated within their own clusters. (As long as one includes babies, IQ scores correlate very well with height.) To this reviewer's mind, Stations 2, 10, and 11 should not have been included in the evaluation of model predictions for the stations on the ridge. Their domination of the correlation coefficient imparts no physical meaning to its interpretation. On the contrary, their inclusion precludes any coherent interpretation of the results.

Using the averages of the "top 25" across the stations, one calculates correlations similar to those produced by TRC. When one removes the three low stations and recalculates the correlation, one finds that the correlations across all of the models are not significantly different from zero, ranging from 0.26 to 0.66. For n=8, the correlation coefficient must be larger than .666 to be significantly different from zero at the 95% confidence level. Five sets of random numbers produced correlations with the station concentrations that ranged from 0.15 to 0.73. The anchoring of the correlations by the 3 low-concentration stations would have been immediately obvious if some graphs had been used. One thing that is not pointed out in the report is that, when n=11 a correlation coefficient that is less than .576 is not significantly different from zero at the 95% confidence level.

Exploratory Analysis of Top 25 at Westvaco

The purpose of this particular section is to present initial elements of the first step of an evaluation in order to draw attention to the limitations of the TRC report. This reviewer considers the type of exploration of the data presented in this section very necessary, not only as an aid to the interpretation of the evaluation statistics, but also for its inclusion as part of the evaluation itself. The data used is not available in the TRC report and had to be obtained as "outside" information. Although the review ostensibly is to be based on the TRC report, the goals of the overall project stated in the TRC report are better served by bringing in "outside" information and including it in the review, as was done with on-site turbulence data.

Table 3 shows a stem-and-leaf diagram for the top 25 concentrations at Westvaco, depicting and comparing observed concentrations and predictions from the eight models. The stems are to the left of the vertical line and are the left-most 1 or 2 significant digits of the concentration values. The very next digit, rounded, to the right in the concentration value becomes the leaf, the number on the right of the verticle line. Each number to the right of the vertical line represents a different concentration value. Thus the stem-and-leaf display is like a histogram, but it retains

Observed

```
25 | 7
25 | 4
24 |
24 |
23 | 66
23 | 4
22 |
22 | 04
21 | 9
21 |
20 |
20 | 1
19 | 59
19 | 4
18 | 6
18 | 4
17 | 557
17 | 0234
16 | 68
16 | 12
```

CPLX I

```
24 |
24 |
23 | 6
23 |
22 |
22 |
21 | 9
21 | 0
20 | 89
20 |
19 | 9
19 | 113
18 | 9
18 | 00
17 | 569
17 |
16 | 9
16 | 00024
15 | 69
15 | 12
```

CPLX II

```
50 | 7
49 |
48 |
47 | 6
46 | 2
45 |
44 |
43 | 6
42 | 2
41 | 6
40 | 22
39 | 0
38 | 24
37 | 88
36 | 577
35 | 359
34 | 079
33 | 577
```

CPLX/PFM

```
34 |
34 | 2
  =≈
28 |
27 |
27 |
26 |
26 |
25 |
25 |
24 | 5
24 |
23 |
23 |
22 |
22 |
21 |
21 | 0
20 | 9
20 |
19 |
19 |
18 |
18 | 0
17 | 8
17 |
16 |
16 | 4
15 |
15 |
14 |
14 | 013
13 | 9
13 | 223
12 | 59
12 | 124
11 | 7
11 | 224
10 | 78
```

RTDM

```
5 | 0
4 |
4 | 7
4 |
4 |
4 | 3
4 | 12
4 |
3 | 9
3 | 7
3 | 5
3 |
3 | 44
3 | 2
3 | 01
2 | 999
2 | 777
2 | 5
2 | 445
2 | 3
```

4141

```
20 |
19 |
18 | 07
17 | 6
16 |
15 |
14 | 25788
13 | 035
12 |
11 | 0167
10 | 7
 9 | 024478
 8 | 07
 7 | 4
```

PLUMES

```
26 | 6
26 |
  =≈
23 |

22 |

21 |

20 | 2

19 |
18 |
18 |
18 |
17 |
17 | 6
17 |
16 | 99
16 |
16 |
15 |
15 | 5
15 | 12
14 | 9
14 |
14 | 1
13 | 7
13 | 46
13 | 02
12 | 9
12 | 4456
12 | 03
11 | 779
```

SHORTZ

```
24 | 2
  =≈
22 |
   | 8
21 |

20 |

19 |

18 | 0
17 |
17 |
17 | 11
16 |
16 |
16 |
15 |
15 | 4
15 | 0
14 |
14 |
14 |
13 |
13 | 4
13 |
12 | 8
12 | 455
12 | 0
11 | 7779
11 | 5
11 | 011
10 |
10 | 4566
```

IMPACT

```
101 | 6
   =≈
 70 | 0
   =≈
 57 | 4
   =≈
 54 | 9
   =≈
 20 | 0
 19 | 6
   =≈
 17 | 6
 17 |
 16 |
 16 |
 15 |
 15 |
 14 |
 14 |
 13 |
 13 |
 12 |
 12 | 0
 11 |
 11 |
 10 |
 10 |
  9 | 8
  9 |
  8 | 9
  8 |
  7 |
  7 | 5
  6 |
  6 | 12
  5 | 79
  5 |
  4 | 779
  4 | 22
  3 | 66699
```

Table 3: Stem-and-Leaf Displays of Top 25 Values, Westvaco Data; the Number of Stems Between Lowest One and the One Including the Median is the Same (Within Reason) Across All Displays.

numerical information about the concentrations. The diagram was set up so that number of stems between the median (n=13) and the lowest value (n=25) is approximately equal for each of the distributions. The distributions are, therefore, "normalized," although the normalization is not a cardinal one but rather with respect to spread.

There are several items of note. First, four of the models have problems with "outliers" when their normalized distributions are compared to that of the observations. Second, except for IMPACT, removing two of the extreme values brings the spread across bins of five of the models very close to each other. The spreads of SHORTZ and COMPLEX/PFM remain somewhat larger than the rest , but are still very similar to them. One would expect this to occur if the distributions are similar in shape, given the type of normalization. IMPACT stands out as having severe problems. The 4141 model has a noticeably shorter distribution of bins compared to the other models. Third, and importantly, several of the normalized distributions look very similar to the observed data (COMPLEX I, COMPLEX II, and RTDM). Overall the differences are much less than expected, based on the TRC information (e.g., the maximum frequency difference computations), and, except for IMPACT, seem to be mostly differences in "magnification" rather than some gross underlying difference in shape between the distributions. This is a very different conclusion than that presented by TRC.

While the stem-and-leaf display gives a good overview of a particular nature, more insight can be obtained with empirical quantile-quantile (EQQ) plots. Empirical quantile-quantile plots are shown for five of the models in Figures 1-5. In EQQ plots one can also discover the existance of outliers, as for example Figures 3 and 5 for IMPACT and SHORTZ. One can also establish that the distributions of the predicted concentrations are rather similar to the distribution of the observed concentrations, because the EQQ plots are not too far from being straight lines for the lowest ranked 20 points. This is important to establish. However, one now notices that COMPLEX I has an odd "hump" in the middle of its distribution (Figure 1), and that the distribution from SHORTZ is piece-wise linear with two segments (Figure

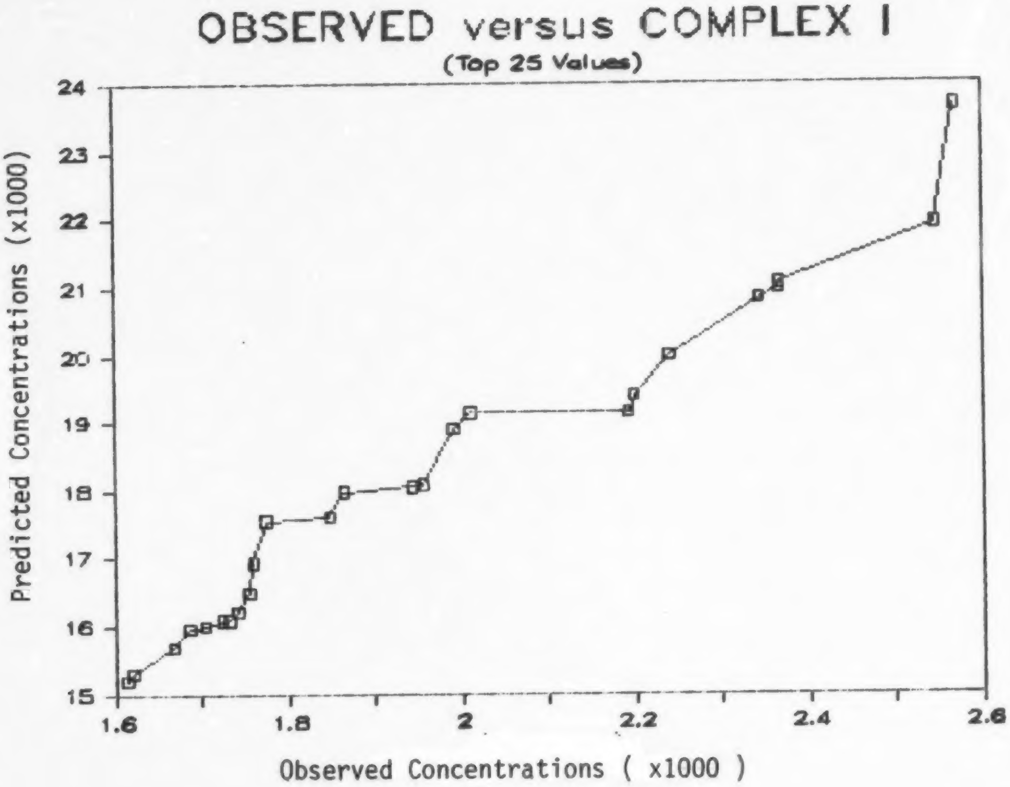


Figure 1: Empirical Quantile-Quantile Plot,
Top 25 Values Westvaco Data

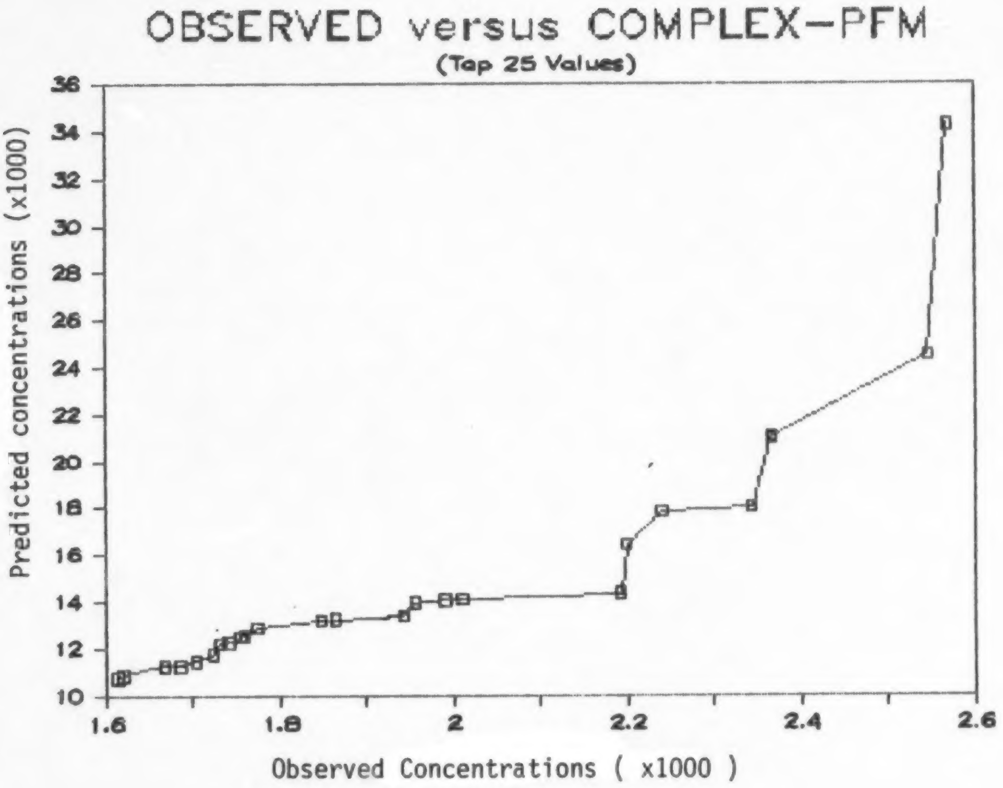


Figure 2: Empirical Quantile-Quantile Plot,
Top 25 Values Westvaco Data

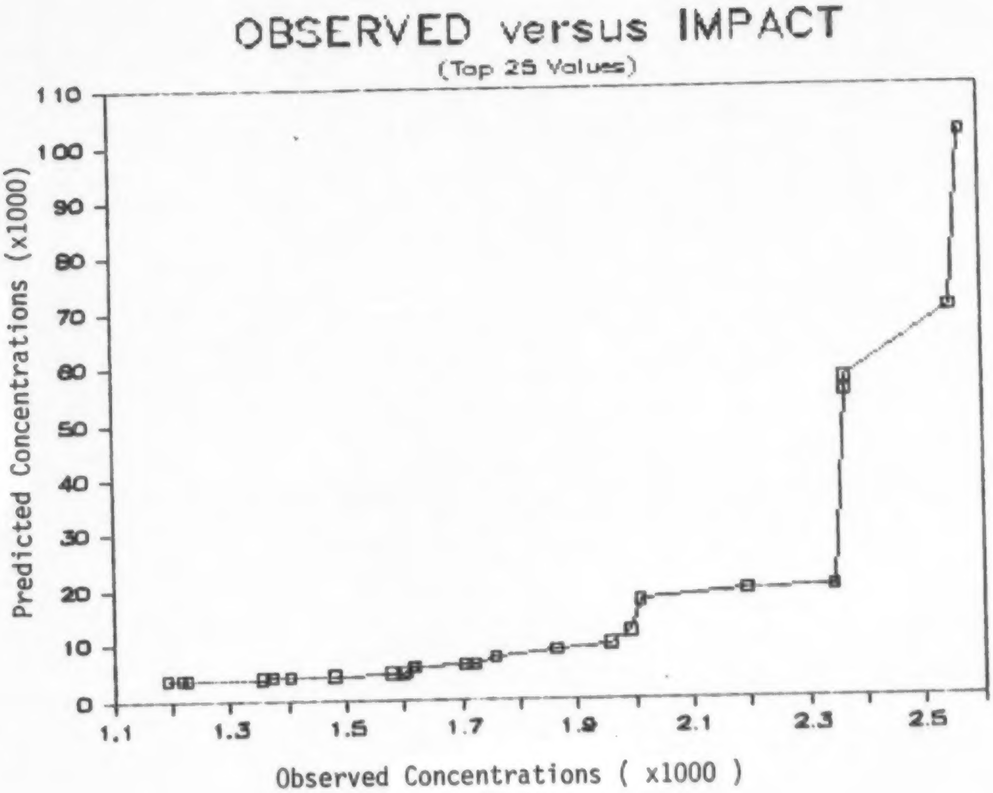


Figure 3: Empirical Quantile-Quantile Plot
Top 25 Values Westvaco Data

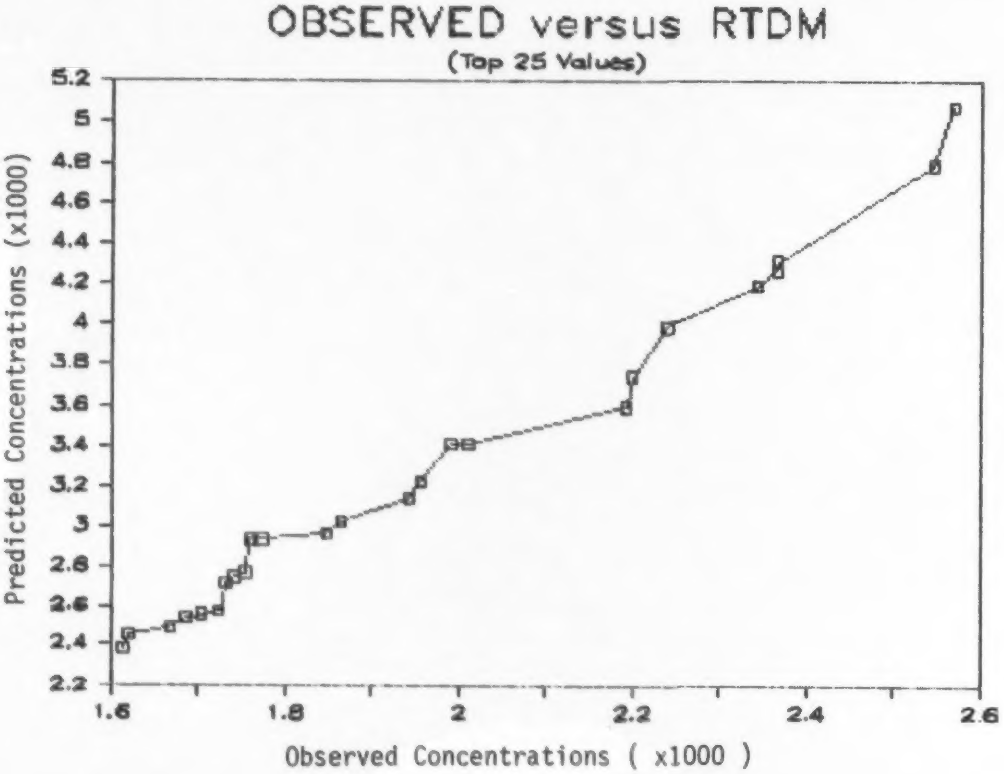


Figure 4: Empirical Quantile-Quantile Plot
Top 25 Values Westvaco Data

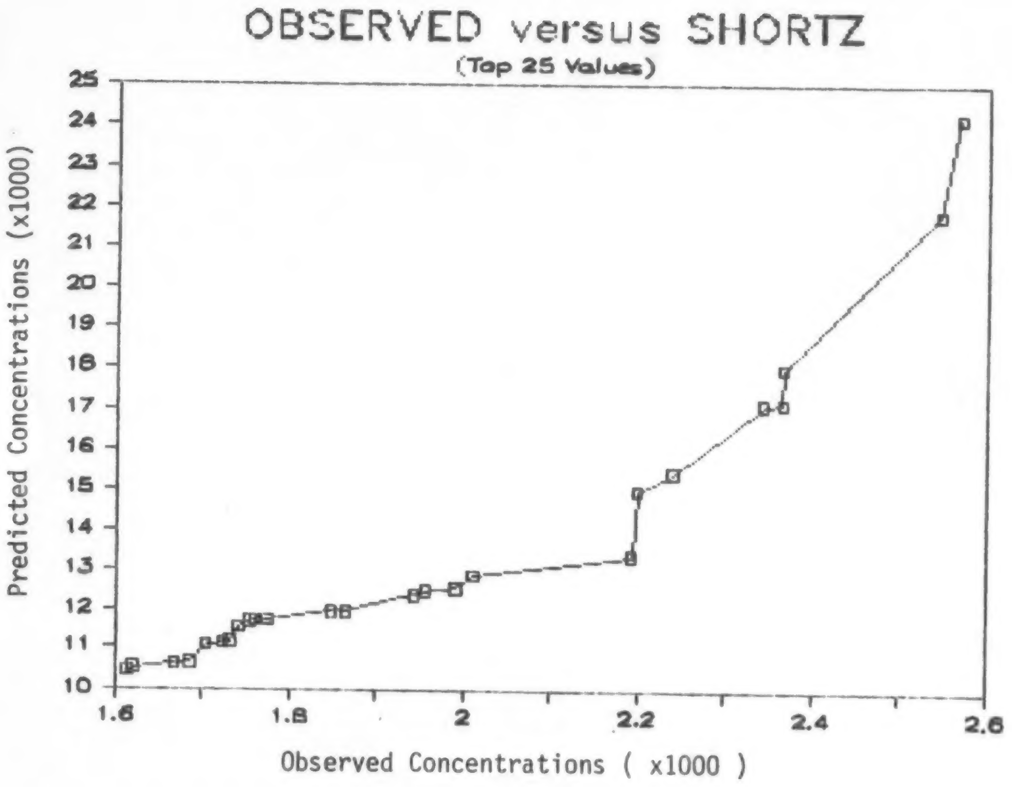


Figure 5: Empirical Quantile-Quantile Plot
Top 25 Values Westvaco Data

5). Thus these two models exhibit odd behavior that is difficult, if not impossible, to notice in the stem-and-leaf display. The severe problems of IMPACT are noticeable in both types of data display.

The EQQ plots establish that the central part of the distribution from each model is reasonably close to being linear with respect to the observed concentrations. Thus one can easily make a quantitative comparison of the spread of the distributions and their "slope" by using the upper and lower quartiles and be satisfied that such comparisons are consistent across all the data sets. Let us again note the need to scale or normalize the distributions for comparison purposes and define a relative spread as the upper quartile minus the lower quartile divided by the median. (We effect a more usual cardinal normalization.) The "slope" we will define as the ratio of the upper quartile to the lower quartile. The results for the models are shown in Table 4.

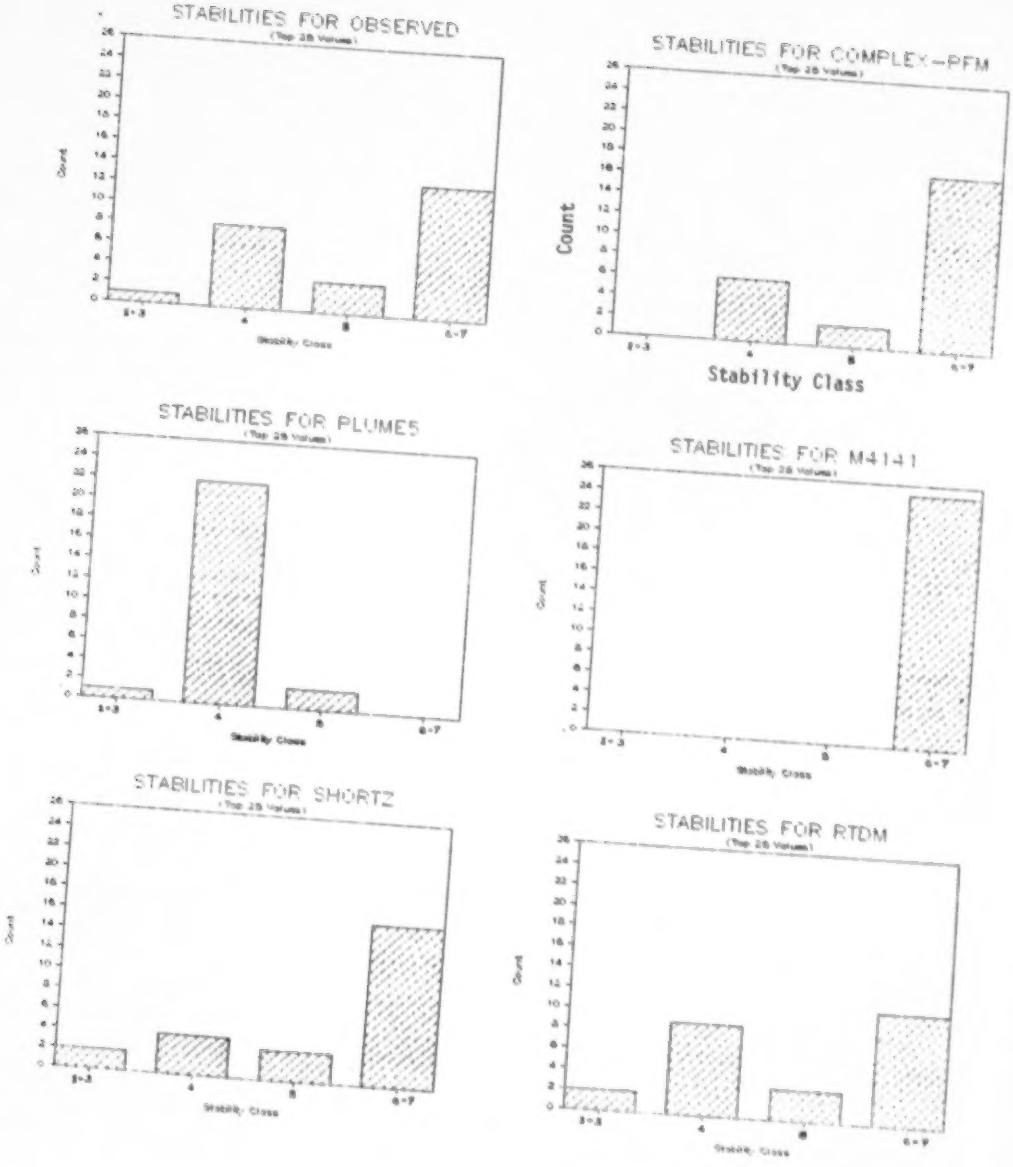
Table 4. Comparison of the Relative Spread of the Concentration Distributions

Model	Relative Spread	"Slope"
COMPLEX I	.200	1.22
COMPLEX II	.154	1.17
4141	.444	1.55
RTDM	.402	1.46
PLUME5	.223	1.24
COMPLEX/PFM	.391	1.43
SHORTZ	.337	1.36
Observed	.264	1.29
IMPACT	2.33	4.37
Observed(IMPACT) (subset)	.387	1.46

The "slope" for SHORTZ is a bit overestimated due to SHORTZ's odd distribution and might be a bit underestimated for PLUME5. But in general, IMPACT is the only model that is clearly out of line with the observations. An uncertainty analysis and statistical test is needed to establish whether or not the spread of a model such as RTDM is significantly different from the observed spread. RTDM and COMPLEX/PFM are rather similar on these measures of relative spread and "slope."

Although some models show reasonable patterns of the "top 25" at Westvaco, ignoring the outliers and the problem of extreme bias, one still would like to know if there is any match with real world conditions associated with the maxima, because the models are going to have to be used elsewhere than Westvaco. We still have to use complete unpairing in time and space, because the models just don't produce maxima on the same days as maxima are observed. The next best match is for the models to reproduce the mix of stability classes, the mix of wind speeds and the mix of stations associated with the "top 25." A first examination of this can be performed with the use of histograms. Figure 6 shows histograms for a breakdown of the top 25 by stability category for five models and the observed. Figure 7 shows histograms of the data broken down into wind speed ranges for five models and the observed. Figure 8 shows histograms for the top 25 data broken down by monitoring sites on the ridge plus station 2 on the other side of the valley.

These three figures show univariate patterns which can be compared with the pattern in the observations. Whereas the range of wind speeds in the top 25 is too narrow for COMPLEX I, it is too broad for RTDM. RTDM and COMPLEX/PFM look the best on stability class comparisons and PLUME5 and 4141 look worst. Most models exhibit the same patchiness in hitting the monitoring stations as demonstrated by the behavior of RTDM. COMPLEX/PFM is the only Gaussian model that moves away from this patchiness. IMPACT looks very good, if it were not for the fact that IMPACT predicts a high concentration for station 2, a prediction which is totally out of line.


STABILITIES FOR OBSERVED

STABILITIES FOR COMPLEX-PFM

STABILITIES FOR PLUMES

STABILITIES FOR M4141

STABILITIES FOR SHORTZ

STABILITIES FOR RTDM

Figure 6: Histogram of Top 25 Values of Westvaco Data: Stability Class

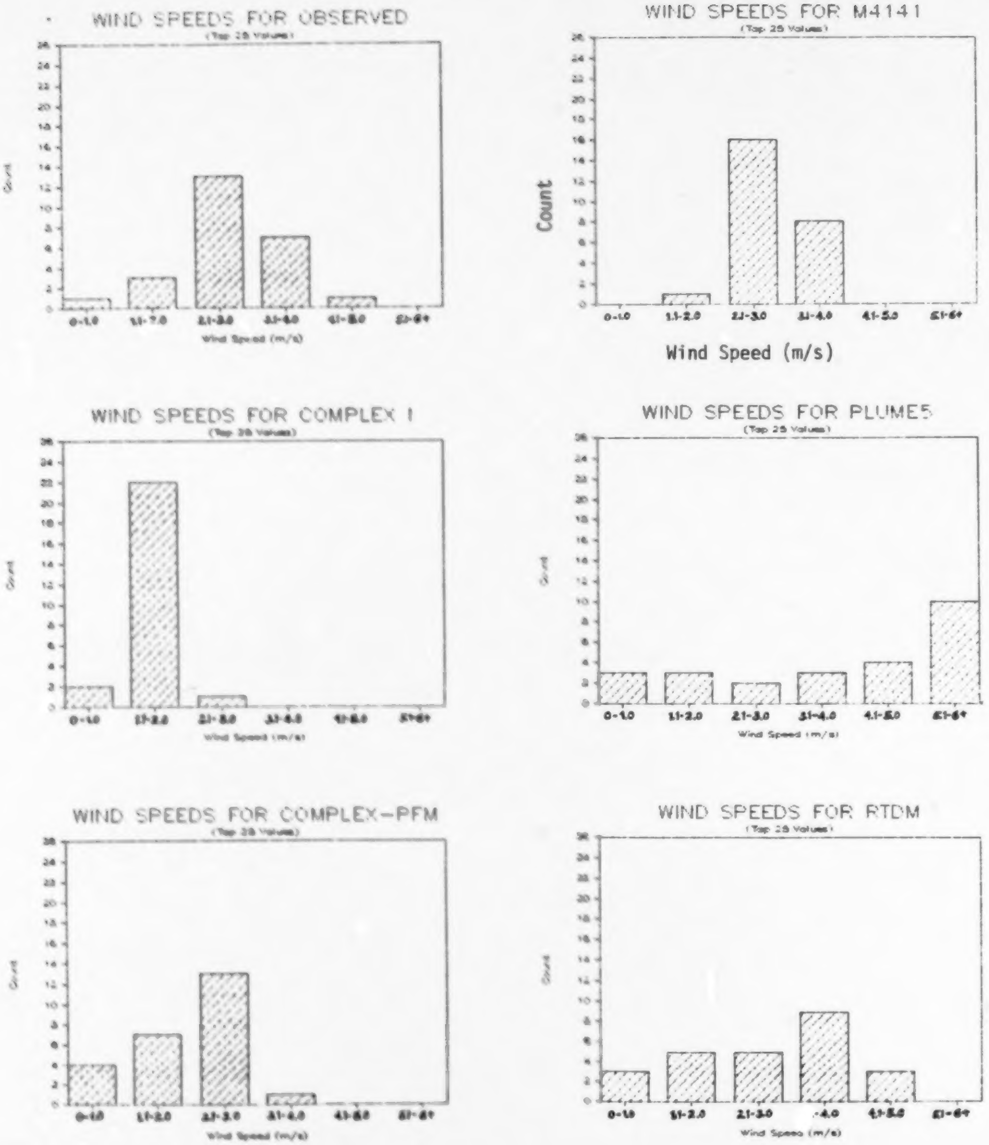


Figure 7: Histogram of Top 25 Values of Westvaco Data: Wind Speed

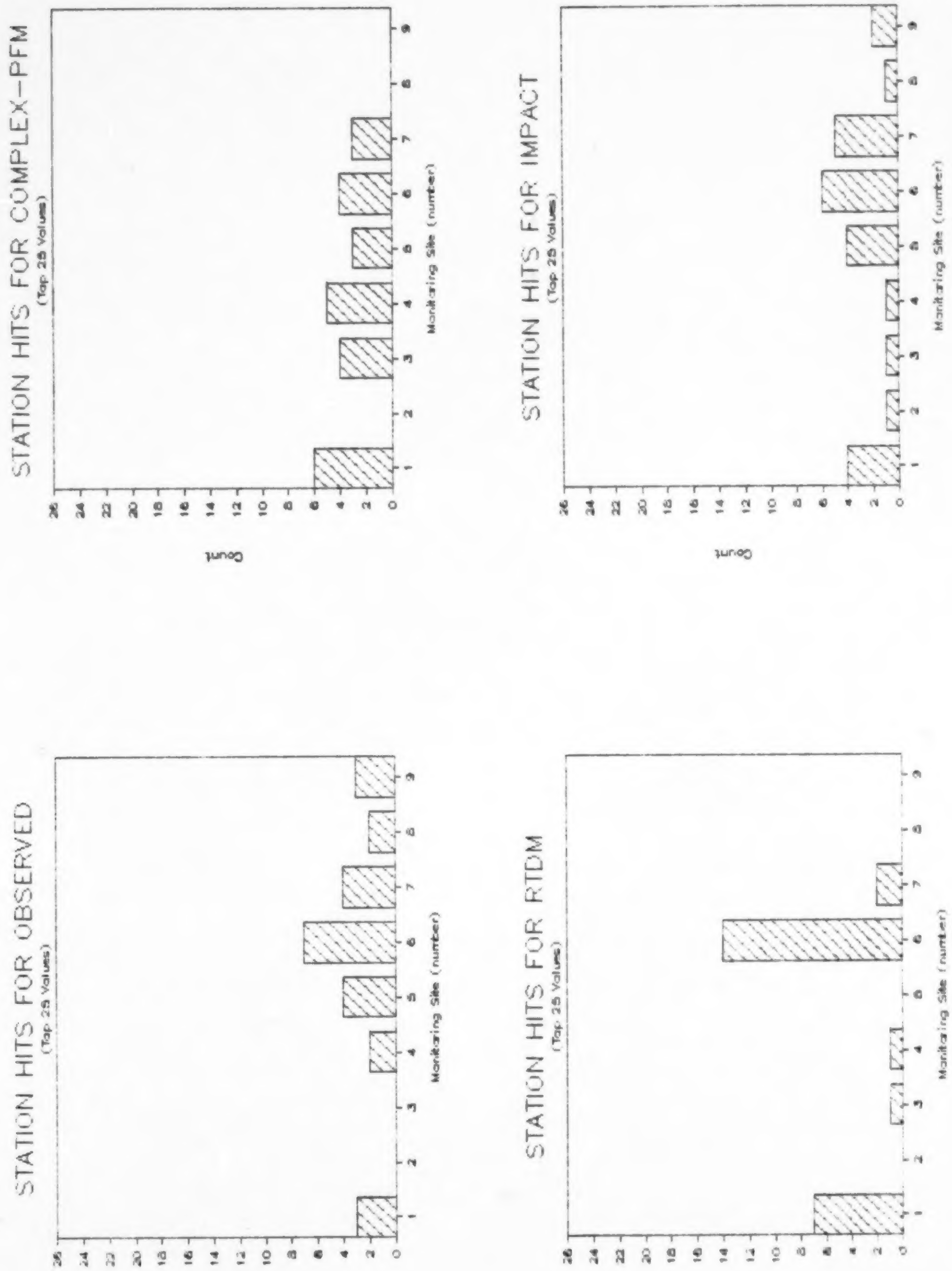
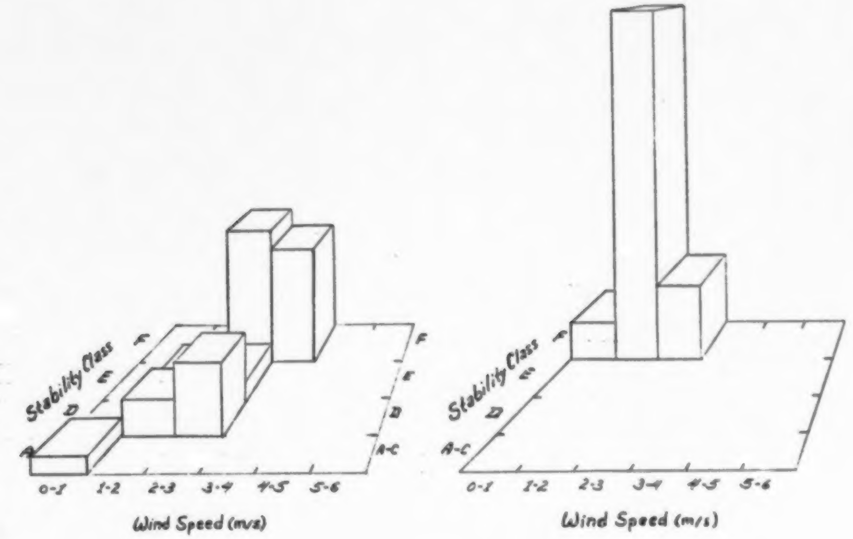
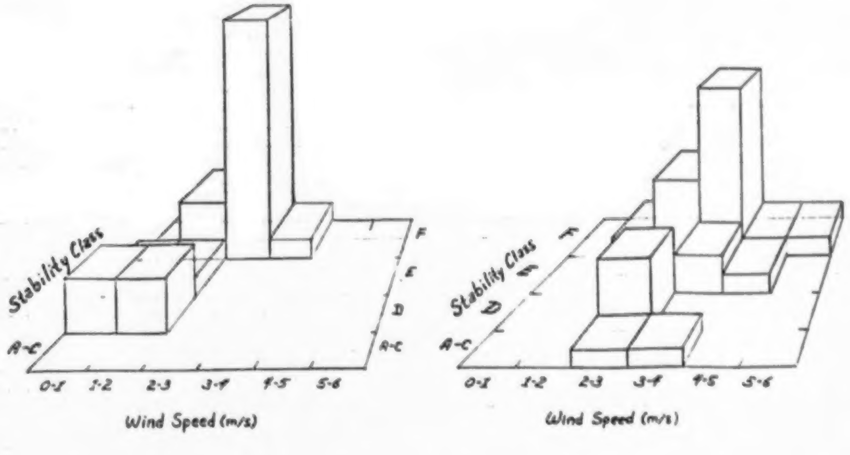


Figure 8: Histogram of Top 25 Values of Westvaco Data: Station Number



Observed **COMPLEX I**



COMPLEX-PFM **SHORTZ**

Figure 9: Bivariate Histogram of Top 25, Westvaco Data, for Combinations of Stability and Wind Speed

Observed

Wind Speed (m/s)

Stability Class	0-1	1-2	2-3	3-4	4-5
1-3	1				
4		2	5	1	
5		1	1		1
6-7			7	6	

[diagonal]

COMPLEX I

Wind Speed (m/s)

	0-1	1-2	2-3	3-4	4-5
1-3					
4					
5					
6-7	2	19	4		

[clump]

COMPLEX/PFM

Stability Class	0-1	1-2	2-3	3-4	4-5
1-3					
4	3	3			
5	1	1			
6-7		3	13	1	

[diagonal]

SHORTZ

	0-1	1-2	2-3	3-4	4-5	5-6
1-3			1	1		
4			4			
5				2	1	
6-7		1	4	9	1	1

[triangle]

Table 5: Counts for the Bivariate Comparison Matrix of Wind Speed and Stability Class.

A more stringent type of match is a bivariate one. Figure 9 shows a bivariate histogram of stability class matched with wind speed versus count. The counts are given in Table 5. Three models are shown. The observations exhibit a diagonal pattern. COMPLEX I exhibits a clumped pattern; COMPLEX/PFM exhibits a diagonal pattern; and SHORTZ exhibits a triangular pattern. This information, the unpaired comparison of model behavior to real world behavior, appears to be useful in helping to discriminate between models.

It should be clear by now that a detailed analysis such as the one begun here is necessary to understand what the models are "doing," and, hence, understand what is contained in the overall statistical measures. Without this understanding one cannot make an informed evaluation of the models on this data set, much less make any informed judgments with respect to the regulatory use of these models. Much more needs to be done. Yet, the TRC document is incapable of providing the necessary information, because what has been used in this subsection is the actual ordered hourly data, not the summaries given in the TRC document.

Some Assessments of Model Behavior (TRC Document)

Almost all of the influences that come from "current" thinking about complex terrain models move the predictions of the models towards lower values compared to the predictions of models based on older concepts and approximations. Getting rid of the surface reflection at impact, using onsite turbulence data, and accounting for terrain shape for ridge-like features would reduce the predictions in most models. Accounting for streamline response to hill-like shapes is the only change that would increase the predictions of the models (according to potential flow theory).

Because all of the models, except IMPACT on Cinder Cone data, suffer from problems of over-prediction, inclusion of "better" science should mean "better" model predictions, according to this evaluation. However, evaluations at two sites, especially in the manner they have been presented, do not make a strong case for such a sweeping conclusion, given all of the uncertainty. Inclusion of a lot more diagnostic information in the evaluation is necessary. A more thorough evaluation might or might not help to bolster the suggestion that better science means better predictions. It is difficult to tell from the information available in the TRC report which way it would go.

One conclusion one might draw from the information presented in the TRC report is that the model that does best, RTDM, has nearly all of its components identified with the more current thinking about how to model complex terrain. But one is not completely comfortable with that conclusion. COMPLEX/PFM, 4141, PLUME5 and SHORTZ are fairly similar in their "top 25" predictions, unpaired in space and time. This seems to be due to very large predictions for a few specific cases of location and stability. How would fixing up the individual components improve these models? One has the "feeling" that they still would not do as well as RTDM.

RTDM seems to be better than the sum of its parts. Why? The TRC evaluation, to this reviewer's mind, is incapable, as it stands, of establishing why. It is not clear whether the needed information from the model predictions can be recovered from what TRC archived to adequately address the above question. But establishing the reasons for RTDM's better performance would be very helpful to the prescreening of other models for evaluation and to the interpretation of their evaluations on the same two data sets as well as other data sets. One consistency test of the reasons underlying RTDM's good performance would be to up-date the two or three major components in COMPLEX/PFM that obviously need it, i.e., use of on-site turbulence data and using IOPT(25) = 5 (or something similar in intent) and rerun the evaluation for this improved version.

Three of the models, PLUME5, 4141, and COMPLEX/PFM seem to have odd behavior that deserve mention. The behavior of PLUME5 does not seem to be consistent with its description. For example, one would expect that for stability classes E and F the predictions from PLUME5 would be larger than those from SHORTZ. They are not. One would also expect that for stabilities D, E, and F, the predictions from PLUME5 should resemble those from COMPLEX II, because the plume rise equations and the diffusion parameters are the same. Again, they do not. There is no reason from the description of PLUME5 to expect that, by a large margin, the maximum concentrations would occur with neutral stability for the Westvaco data. Impaction of the plume is supposedly allowed and expected under stable conditions and a modified half-height correction for neutral stabilities precludes impaction for that case.

The behavior of 4141 does not seem to be very consistent on the Westvaco data. It too has markedly larger predictions for a single stability class. However, it is 4141's spatial behavior that is more unusual. This can best be seen by looking at the relative rankings of the "top 25" model predictions as a function of distance from the source, using 1-hour concentrations as shown in Table 6.

Table 6. Ranking of the Models Based on the Highest 25 Values by Station Number

	Monitoring Station			
RANK	1 (800m)	4 (900m)	5 (1100m)	9 (1500m)
1	CPLX II	CPLX II	CPLX II	4141
2	CPLX I	CPLX I	CPLX I	CPLX II
3	CPLX/PFM	SHORTZ	4141	CPLX I
4	SHORTZ	CPLX/PFM	CPLX/PFM	CPLX/PFM
5	PLUME5	4141	RTDM	RTDM
6	4141	RTDM	PLUME5	SHORTZ
7	RTDM	PLUME5	SHORTZ	PLUME5

Whereas the models by-and-large maintain their rank ordering as a function of space, 4141 does not.

COMPLEX/PFM does not seem to be operating in accordance with its description. For unstable conditions the predictions from COMPLEX/PFM should be the same as COMPLEX II. They are not. The code indicates that COMPLEX II is being called for the correct stability classes. For neutral stability one would expect the predictions from COMPLEX/PFM to be less than those from COMPLEX II for a ridge and greater for a hill. In the Westvaco data there is a tremendous difference. One can only surmise that this difference is due to differences in plume rise or that something is drastically wrong. For Cinder Cone, the predictions from the two models (top 30) are essentially equal. This comparison is marred by the fact the C stability is mixed in and the comparison is not for a pure neutral case. For cases with a stable atmosphere, one would expect predictions from COMPLEX/PFM to at least be less than or equal to predictions from COMPLEX I. For Cinder Cone, where plume rise is not a complicating factor, this is not the case.

One model, IMPACT, can not truly be evaluated because there is no equivalent information to that in Table 5-2 of the TRC report for the data set that includes its predictions. This is very unfortunate and should be rectified. Otherwise a tremendous amount of effort by TRC is wasted and a considerable amount of useful information is lost. The behavior of IMPACT does appear to be inconsistent. It had the lowest predictions for Cinder Cone and some of the highest for Westvaco. In fact, on the 10-station average of the highest concentrations predicted, IMPACT predicted concentrations more than twice those of COMPLEX II (45,000 versus 19,000). Thus IMPACT reverses its ranking across the two data sets, as indicated below.

Table 7. Ranking of Models Based on the 25 Highest Values (unpaired in time and space)

RANK	Westvaco	Cinder Cone
1	COMPLEX II	COMPLEX II
2	IMPACT	4141
3	COMPLEX I	PLUME5
4	COMPLEX/PFM	COMPLEX/PFM
5	SHORTZ	SHORTZ
6	PLUME5	COMPLEX I
7	RTDM	RTDM
8	4141	IMPACT

One notes that, in fact, IMPACT changes places with 4141 and PLUME5 changes places with COMPLEX I. Although based on admittedly weak evidence, one slowly begins to build an impression that the predictions from IMPACT, 4141, PLUME5 and possibly COMPLEX I should not be trusted.

Assessments Revisited

Did the information contained in the more detailed exploration of the top 25, based on the actual data, make a significant contribution to an assessment of the models? Clearly, the answer is yes. The EQQ plots showed that for the top 25 at Westvaco the differences in sample populations of the predictions does not seem to present a problem for the comparisons. The EQQ plots showed that SHORTZ is the only model that has a problem in this regard. This gives notice that something is different in the predictions of SHORTZ. Except for SHORTZ and except for extreme outliers, the distributions from the Gaussian models were differing mostly by a factor of "magnification," which was severely distorting the interpretation of the statistical measures used by TRC. IMPACT has severe problems with outliers.

A much worse problem that this exploration pointed out seems to be one of outliers. The outliers do affect the statistics in important ways. For example, trimming the top three outliers changes the mean of the top 25 predictions from IMPACT from 17,829 to 9,845; a substantial change. The models do not exhibit similar behavior with respect to outliers: some have them, others do not; of the model which had outliers some had a few, one had many. The stem-and-leaf display and the EQQ plots are important tools to use in defining procedures to deal with this problem of outliers, such as trimming the distribution.

The EQQ plots established the fact that the upper and lower quartiles could be effectively used to quantitatively compare the spreads and the "slopes" of the distributions from the different models. With appropriate trimming, which will be case specific, one can now go back and compute statistics that could be used for hypothesis testing of the differences or perform robust regression on the distributions to compare slopes. The comparisons of the distributions based on the information in the TRC document were totally inadequate as well as being misleading.

Importantly, the new information changed this reviewer's mind about the ability of better science to improve the predictions for regulatory applications of the models. Some worrisome, odd behavior of SHORTZ began to surface, whereas, before, its predictions seemed not that different from COMPLEX/PFM's. COMPLEX/PFM is doing better than the other models in more consistently having a pattern of prediction somewhat like the real world pattern. The older models are more consistently locked-in to categories of wind speed and stability class that did not match well with the observations, engendering low confidence in their predictions for other situations. COMPLEX/PFM has many similarities with RTDM. That is encouraging. Thus, maybe even RTDM could be improved by putting in better science. It certainly looks worthwhile to upgrade COMPLEX/PFM and get it away from the older model formulations, especially away from COMPLEX I, and get it away from using VALLEY-like computations when there is impaction of the plume. The station comparison for IMPACT indicates that possibly there are advantages to working with three-dimensional wind fields, if the other problems can be corrected.

Thus, to this reviewer's mind, the use of graphs and pattern comparisons and other techniques of exploratory data analysis is very important to the interpretation of model performance and the interpretation of the aggregate statistical measures, as used by TRC. More should obviously be done than presented in this review, because this paper is a review, not an evaluation. One should be able to generate such analyses; however, one cannot from the information presented in the TRC document. In a sense, the document is remiss in meeting the stated objectives. The document, as it stands, cannot support a performance evaluation of the models. At a minimum, the raw data need to be provided for each class of breakdown of the data.

REVIEW OF THE EVALUATION FRAMEWORK AND PRESENTATION

This section will address three sets of issues: First, there are issues with respect to the larger question of the representativeness of an evaluation based on the two data sets. Second, there are issues that need to be addressed with respect to how the measures are developed and used. Third, there are the more minor issues of useful and complete presentation of the information.

Questions of Representativeness

The stated goal of the TRC evaluation is: " The principal objective of this project is to produce performance statistics so that EPA and a group of reviewers may judge the relative merits of different models." The judgments are desired for the purposes of "... a systematic evaluation of these models to decide in an objective manner which models should be included in the guidelines and what recommendations should be made concerning the use of these dispersion models for regulatory application."

No evaluation data set can meet every need of an evaluation. That is obvious. The point is how to make the best use of the data. An important step in that direction is to clearly define the representativeness (or limitations) of the data set with respect to the goals of the evaluation. This is true, whether or not prior thought has gone into structuring and designing the data set from the point of view of a model evaluation. Clearly a lot of thought went into the Cinder Cone experiment.

Both data sets test the "close-in" behavior of models. For the Westvaco case, plume-rise calculations seem to have a very important influence. Thus the Westvaco data set could be identified as much with tests of transient (close-in) plume-rise behavior as with complex terrain behavior. In this evaluation, thought does not seem to have been given to dealing with the effect of plume rise on the evaluation results, and trying to more systematically account for it.

Close-in behavior also means short travel times, on the order of minutes, even for 1 m/s winds. Both data sets test the value of using on-site turbulence data for a regime that is similar to that used for the development of the PGT curves. This is excellent. But the spatial regime of one kilometer is quite limited compared to the full range of distances to which the models are expected to be applied.

Cinder Cone was obviously an excellent choice to represent a simple hill feature. Westvaco is not the best representative of a two-dimensional terrain feature, but the history behind the choice of Westvaco is not known.

What about applications of the models for longer distances? Many applications are cases in which the receptors are at a much greater distance from the source than for this evaluation; past the point of final plume rise, beyond a 1-hour travel time from the source, and influenced by upwind topographic relief. How important is it to use on-site turbulence measurements for these cases? Over what distances are these measurements valid? Insights from the Westvaco and Cinder Cone evaluations will not be directly applicable to answer such questions.

Also, model behavior can be a function of distance. For example, for stable conditions and impaction of the plume, COMPLEX I predicts values that are one-half those of COMPLEX II at a one-kilometer distance. At five-kilometers distance, COMPLEX I predictions are one-fifth those of COMPLEX II. Assuming the other models do not greatly change their performance relative to each other and RTDM doesn't do something odd, then at five kilometers, COMPLEX I could possibly be in second place, right ahead of COMPLEX/PFM, as far as the top 25 values are concerned (assuming overprediction is still a problem at five kilometers). Both COMPLEX I and COMPLEX/PFM will produce better predictions than the other models (except RTDM), but not because they do a better job of treating complex terrain. At some distance they may do a "better job" than RTDM. This points out that one should be suspicious of taking the results of a performance evaluation at one distance and "blindly" using those results as the basis from which to judge the behavior of the models at all distances.

Thus one must think about the issues central to the evaluation of a complex terrain model and central to the operational use such a model will be put to. From a detailed understanding of these issues, one can develop a number of criteria that can help guide the establishment of a number of evaluation data sets. Sensitivity tests of the models could also be required as part of the evaluation. The present evaluation can serve an important function with respect to learning how to perform better evaluations of complex terrain models. There is no substitute to going through an actual evaluation. A good start has been made, but more work has yet to be done. More work has to be done in defining what is needed of performance evaluations to adequately address the broad goals stated in the TRC document. One obtains the impression from this evaluation that the issues have not been thoroughly enough thought through.

Questions About the Approach to the Measures

One has the distinct impression that the measures used in this evaluation have been implemented in a very mechanical fashion. There is no argument with the issues the measures are supposed to address; the issues are good ones. The argument is that, while there may have been a lot of thought devoted to the development of a potentially useful measure, very little time or thought seems to have gone into making sure the measures are actually giving us relevant information, doing the job they are supposed to, once they are actually applied to a real "live" evaluation. Two examples of severe problems with the interpretability of measures were presented above: the measures comparing frequency distributions and those of spatial correlation (Pearson and Spearman correlation coefficients). It seems clear that these measures were mechanically computed without spending much time to think about what was being computed and what, if anything, could be influencing the answers and/or severely distorting them.

Another example along the same line might also be instructive. This example has to do with sensitivity to extremes. One of the issues of model evaluation has been that the accuracy of highest or second highest estimates from a model is expected to be poor. Guidance from the scientific community has been that evaluations applied to an upper percentile of the predicted values would be more informative about overall model performance than those applied only to the extremes (Fox, 1981).

First, from an examination of the numbers in the TRC report, one suspects that even though the authors followed the recommendations in Fox (1981) and used the "top 25" (the 2.8th percentile), the top 25 for some models, particularly IMPACT and COMPLEX/PFM, appear still to be affected by a few very extremely high predicted values. One must recompute the average of the maxima for the stations at Westvaco on the basis of the stations on the ridge (n=8) to see this more easily. The problem with outliers is immediately obvious from the stem-and-leaf and EQQ plots, however. Thus the intent of the recommendation about extreme estimates is most likely not being met in this evaluation. (The problem with the predictions from IMPACT may be an extreme case.) No check of the behavior of the extremes seems to have been done to look for values that are truly "wild."

The point is that other approaches exist which could easily address the problem just mentioned. It seems to be quite valid to apply the concept of "trimmed means" to these evaluations (Mosteller and Tukey, 1977). That is, the sample is trimmed of its possibly straggling tails by setting aside some fraction of the values from each tail of the sample. Considering our special case, the trimming could be asymmetric, only the upper side of the distribution would be trimmed. This would result in a mean that would be less sensitive to the extremes, yet would well (possibly better) characterize the behavior of the upper distribution. The other approach is to use the median of the population, rather than the mean. Each approach has its advantages and disadvantages, but both seem to be better than what is presently done in this evaluation. Both of these suggested approaches would produce answers which are closer in spirit to the intended AMS recommendation.

Median differences were reported by TRC, but it seems like they were not integrated into the reporting of results.

Since the extreme estimates of the models are still going to be used in regulatory practice, it seems important that the behavior of the models for the extremes be explicitely examined as well. The extremes should not be ignored or hidden in the upper percentile just because they are difficult to predict. Thus it would be useful to evaluate both the trimmed means and the extremes that were trimmed. This raises some important issues about the potential difference between an evaluation and an application. On what basis is one to judge the model?

Second, for the different models, the "top 25" means different things, that is, this set of values is comprised of different sample populations. For some models, the "top 25" means a sample of the top 25 of a particluar population. However, for other models, the "top 25" means a sample of the top 2 to 4 from several populations. Thus some "top 25"'s have more extreme values (top 2-5 from several distributions) than others. As a result, some models are being evaluated with a large weight on their extreme predictions, whereas other models are not having such a large weight put on their extreme predictions. This would seem to lead to some inconsistencies and problems with the comparison between observations and prediction and with cross-comparisons of the model predictions. The question that needs to be addressed is, does the existence of this inconsistency make a difference? It turns out that, except for SHORTZ, the inconsistency probably does not make a difference for the Westvaco data. This was empirically demonstrated with the EQQ plots. What about Cinder Cone? Not enough attention has been given in this evaluation to the issues involved in unpairing in time and space and what haphazard unpairing is doing to the evaluation. This area would seem to deserve some attention.

"Informative graphic techniques should be included in any performance evaluation." (Fox, 1981, p.603). Graphics are noticeable by their total absence in this evaluation. That is remiss. Some of the issues about distributions raised above can be addressed very effectively through the use of graphs, as we have seen in this review. Graedel and Kleiner (1983) discuss several types of graphs that could be useful in performance evaluations, including the empirical quantile-quantile plots. The usefulness of this evaluation has been limited by the exclusion of graphs.

General Presentational Details

I have a few specific comments concerning the TRC report, mostly with respect to fairly obvious suggestions about the text or the tables. The documentation of the models is not considered to be precise enough, although on the whole TRC did a very reasonable job. For example, from the TRC description one has no idea that in 4141 the lateral diffusion parameters of PGT are multiplied by a factor of 1.82. Also, stating that a model uses a modified half-height plume correction does not provide enough information. It would be useful to also state in which direction the modification affects the answer compared to an unmodified half-height correction.

Stations 10 and 11 of the Westvaco study are not defined, either in the text or on Figure 3-3 of the TRC report The definition of the geometry of the monitoring stations relative to the release heights would be very helpful. The labeling of Table 5-2 is not precisely correct for all of its subparts.

In Table 5-5 it would be much superior to give the average predicted value and put the average observed in as a footnote, since the observed is essentially the same for every model (only one exception). In many of the tables only bias is given; the predicted values should also be presented. In all of the tables listing predicted and observed averages, the standard deviations should also be presented. If the data are skewed, as one might expect, then it would be better to report the median and the upper and lower quartiles. That would give some indication of skewness as well as give the magnitude of the interquartile separation, which would be useful for intercomparisons. In tables such as Table 5-7 it would be very useful to include the average predicted value. This is simple and easy to do and can contain useful information. These numbers would have provided some helpful cross-checks in this evaluation.

There were several instances where the Pearson Correlation Coefficient was not significantly different from zero at the 95% confidence level, for example, in Table 5-5. These cases should always be noted. This author tends to prefer gross error over root mean square error, because the latter emphasises the extremes. I would rather have a measure that provides one with a better sense of the central tendency, as does gross error.

It would be useful to have more summary tables of the model predictions in the report. These would give the reader a better overview. Two examples of summary tables are given early in this review. Other summary tables would be ones for each of the breakdowns used, such as wind speed or monitoring station for the Westvaco study.

Finally, a reiteration of points made earlier. One recommendation is that the measures of variance comparison and maximum frequency difference be corrected for bias. The Pearson and Spearman correlation coefficients, as calculated by TRC, are probably not meaningful measures and could be omitted, at least for the Westvaco date. Graphs would be better. Values for C stability should be separated from those of D stability for the Cinder Cone data. It would be very useful to rerun COMPLEX II and COMPLEX/PFM with IOPT(25)=5. A very strong recommendation is that the actual hourly data with all of the various conditions of windspeed, stability class, station number, etc., for each of the breakdowns (subgroupings such as the top-25 or the top-25 by stability class) in the tables developed by TRC should be made available as an appendix to the report. Thought should be given to doing the same for the 3-hourly and the 24-hour values as well.

REFERENCES

Argonne National Laboratory, J.J. Roberts, ed., Report to the U.S. EPA of The Specialists' Conference on the EPA Modeling Guidelines, Environmental Protection Agency, Research Triangle Park, NC, 1977.

Briggs, G.A., Plume Rise and Buoyancy Effects in: Atmospheric Science and Power Production, Darryl Randerson, ed. DOE/TIC-27601 (DE84005177), U.S. Department of Energy, Technical Information Center, Oak Ridge, Tenn., 1984. pp. 327-366.

Deardorff, J.W., "Different approaches toward predicting pollutant dispersion in the boundary layer, and their advantages and disadvantages," WMO Symposium on Boundary Layer Physics Applied to Specific Problems of Air Pollution, Norrkoping, June 19-23, 1978, WMO-No. 510, Geneva, Switzerland, 1978, pp. I.1-I.8.

Dennis, R.L., and J.S., Irwin, Current Views of Model Performance Evaluation, in: Proceedings of the DOE/AMS Model Evaluation Workshop (Oct. 23-26, 1984 at Kiawah, S.C.), Vol. I: Participants and Invited Speakers Papers, A.H. Weber and A.J. Garret, eds, E.I. DuPont de Nemours and Co., Savannah River Laboratory, Aiken, S.C., 29808, 1985.

Egan, B.A., Turbulent diffusion in complex terrain, in: Lectures on Air Pollution and Environmental Impact Analyses, D. Haugen, ed. American Meteorological Society, Boston, MA, 1975, pp. 112-135.

Fox, D.G., "Judging air quality model performance," Bulletin American Meteorological Society, 62: 599-609, 1981.

Gibbons, J.D., Nonparametric Methods for Quantitative Analysis, Holt, Rinehart and Winston, New York, 1976. 463 pp.

Graedel T.E. and B. Kleiner, Exploratory analysis of atmospheric data, in: Probability, Statistics, and Decision Making in the Atmospheric Sciences, A.H. Murphy and R.W. Katz, eds., Westview Press, Boulder, CO, 1983.

Hunt, J.C.R. and P.J. Mulhearn, Turbulent dispersion from sources near two-dimensional obstacles, J. Fluid Mech., 61: 245-274, 1973.

Hunt, J.C.R., J.S. Puttock and W.H. Snyder, Turbulent diffusion from a Point source in stratified and neutral flows around a three-dimensional hill--Part I. Diffusion equation analysis, Atmospheric Environment, 13: 1227-1239, 1979.

Hunt, J.C.R., R.E. Britter and J.S. Puttock, Mathematical models of dispersion of air pollution around buildings and hills, in: IMA Conference on Mathematical Modelling of Turbulent Diffusion in the Environment, C.J. Harris, ed. Academic Press, New York, 1979. pp 145-200.

Irwin, J.S., Estimating plume dispersion--a comparison of several sigma schemes, J. of Climate and Applied Meteor., 22: 92-114, 1983.

Lamb, R.G. and D.L. Durran, Eddy diffusivities derived from a numerical model of the convective boundary layer, Il Nuovo Cimento, 1: 1-17, 1978.

Mosteller F. and J.W. Tukey, Data Analysis and Regression, Addison-Wesley Publishing Co., Reading, MA, 1977. 588 pp.

Pasquill, F., Atmospheric Dispersion Parameters in Gaussian Plume Modeling Part II, U.S. EPA, EPA-600/4-76-030b, Env. Monitoring Series, U.S. EPA, Research Triangle Park, NC, 1976. 44 pp.

Snyder, W.H., R. Britter and J.C.R. Hunt, "A fluid modeling study of the flow structure and plume impingement on a three-dimensional hill in stably stratified flow," in: J.E. Cermak, ed. Wind Engineering, pp. 319-329. Pergamon Press, New York, 1980.

Snyder, W.H., Fluid modeling of terrain aerodynamics and plume dispersion, paper in the Sixth Symposium on Turbulence and Diffusion, Preprint volume, American Meteorological Society, Boston MA, 1983. pp. 317-320.

Wackter, D.J. and R.J., Londgergan, Evaluation of Complex Terrain Air Quality Simulation Models, EPA-450/4-84-017, Office of Air Quality Planning and Standards, EPA, Research Triangle Park, NC 27711, 1984. 243 pp.

REVIEW OF COMPLEX TERRAIN MODELS

Prepared for the AMS-EPA Steering Committee

by

William H. Snyder*
Meteorology and Assessment Division
Environmental Sciences Research Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711

July 1984

---

*On assignment from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce.

REVIEW OF COMPLEX TERRAIN MODELS

INTRODUCTION

Eight complex terrain models are reviewed relative to their scientific merits and their performance measures. Seven of the models are based on Gaussian plume assumptions and one is a numerical grid model. This review will attempt to assess the scientific merits of the various models and explain their performance in terms of those merits.

ASSESSING THE SCIENTIFIC FOUNDATIONS OF THE MODELS

Description of the Models

The seven Gaussian models have basically similar model components. Five models use the wind speed at release height as input whereas two (RTDM and SHORTZ) use a wind speed extrapolated from release height to plume height. Four models use the Turner stability categories to obtain the Pasquill-Gifford sigmas (except for COMPLEX I, which uses 22.5° sector averaging for horizontal dispersion, and COMPLEX/PFM, which uses different combinations of Pasquill-Gifford-Turner (PGT) sigmas and sector averaging for different stability categories). Two (RTDM and SHORTZ) use onsite turbulence data to determine dispersion coefficients. One (PLUME5) obtains stability categories from horizontal turbulence intensities and time of day, then uses Pasquill-Gifford sigmas. All seven models include buoyancy-induced vertical dispersion.

Limits to vertical mixing are identical in five models and very similar in a sixth model (SHORTZ). RTDM, however, uses a reflection factor for terrain that is a function of the slope of the terrain; for flat terrain, this method defaults to full reflection as is customarily assumed, but for sloping terrain, it permits partial plume reflection.

Larger differences appear in the Gaussian models in their calculation of plume rise. COMPLEX I and II use Briggs' (1975) final plume rise, including momentum. PLUME5 uses Briggs' final rise with a determination of stable layer penetration. SHORTZ uses a modification of Briggs' (1971, 1972) final rise with hourly temperature gradients for plumes rising under stable conditions. 4141 uses Briggs' (1975) transitional rise. RTDM also uses Briggs' transitional rise, but with hourly temperature gradients in stable conditions. Finally, COMPLEX/PFM uses a modification of Briggs' layered rise with allowances for reductions in plume height due to streamline deformations. Five of the models allow for stack tip downwash, while two (4141 and PLUME5) do not.

Major differences in the models appear in the manner in which terrain impaction is treated. COMPLEX I and II use terrain adjustment factors of 0.5 (plume half-height assumption) for neutral and unstable flows and zero (with a standoff distance of 10 m) for stable flows. COMPLEX/PFM performs the same as COMPLEX I and II except in the narrow range of stable conditions when the plume is above the dividing-streamline height. Under these conditions, the plume centerline follows calculated streamlines with allowances being made for stability (Froude number) and hill shape (crosswind aspect ratio). 4141 uses the half-height assumption for neutral and unstable conditions and a quarter-height assumption for stable conditions. PLUME5 uses a variant of the half-height assumption. SHORTZ assumes plume impingement within the mixed layer. RTDM uses the half-height assumption for all neutral and unstable conditions and also for stable conditions when the plume height exceeds the dividing-streamline height. It assumes impingement when the plume is below the dividing-streamline height.

The three-dimensional grid model, IMPACT, interpolates and extrapolates wind measurements at multiple sites to create a divergence-free wind field. It uses diffusivities derived from Smith and Howard's (1972) empirical formulations to obtain finite difference solutions to the diffusion equation. The plume path is determined through the computed wind fields, with Briggs' layered plume rise including penetration of stable layers.

Evaluation of the Models

Probably the most crucial question concerning prediction of concentrations in complex terrain is whether the plume impinges on the terrain or surmounts the hill top. Plumes are allowed to impinge on terrain in all models except 4141 and PLUME5. However, the conditions under which plume impaction is allowed to occur are generally much too broad and, under impingement conditions, it would appear that most of the treatments overestimate surface concentrations.

From laboratory studies (Hunt and Snyder, 1980; Snyder et al., 1984) and field studies (Lavery et al., 1982), we know that plume impaction can and does occur. These studies have shown that plume impingement occurs only under strongly stable conditions. The criterion for impingement should be based on a Froude number based upon the approach flow wind speed, the temperature gradient from the base to the top of the hill, and the hill height; more appropriately, the criterion should be based upon the dividing-streamline height (Snyder et al., 1985), which allows for arbitrary shapes of wind and temperature profiles. The assumption of plume impingement under all stable conditions is simply incorrect. Hence, COMPLEX I and II can be expected to predict plume impingement much too often.

Under impingement conditions, COMPLEX II simply calculates surface concentrations with a bivariate Gaussian distribution and a

10 m standoff distance, thus allowing for no terrain effects on plume diffusion and no wind meander. Hence, we may expect COMPLEX II to grossly overestimate surface concentrations because (1) it predicts impingement conditions much too often and (2) under impingement, it does not allow for plume meander, deformation, nor increased diffusion.

COMPLEX I will, in general, predict lower concentrations than does COMPLEX II because it performs a 22.5° sector averaging for horizontal dispersion. However, laboratory studies (Snyder and Lawson, 1981) and field experiments (Strimaitis et al., 1983) suggest that such sector averaging is inappropriate because of (2) above. Hence, COMPLEX I may also be expected to grossly overestimate surface concentrations.

COMPLEX/PFM moves one step in the right direction by determining the hill Froude number and thus is more likely to predict the proper conditions for plume impingement. However, under impingement conditions, COMPLEX/PFM calculates concentrations in the same manner as does COMPLEX I, i.e., 22.5° sector horizontal averaging. Hence, COMPLEX/PFM may also be expected to overestimate concentrations under impingement conditions because of (2) above.

SHORTZ treats plume impaction from an entirely different point of view than do the other models. In SHORTZ, a plume must be contained within a surface mixing layer if it is to cause significant ground-level concentrations at any point. Hence, plume impaction, in the SHORTZ definition, may occur under any stability condition but, in the now customary definition of plume impaction, will not occur under stable conditions unless a surface mixing layer exists that is deep enough to include the plume. As a practical matter, however, the depth of the surface mixing layer is defined as the height at which the vertical turbulence intensity drops below 0.01, so that plumes in complex terrain are almost always within the surface mixing layer (see later discussion).

SHORTZ uses onsite turbulence data to obtain dispersion coefficients, which is certainly an improvement over the PGT scheme, and so SHORTZ may be expected to perform better relative to, say, COMPLEX I or II. Also, because the model has been tested and developed with the use of numerous data sets, we may expect its performance overall to be relatively better than the COMPLEX (I,II, PFM) models.

RTDM uses the dividing-streamline concept to predict plume impaction, although it uses a bulk parameterization (Froude number) as opposed to the more refined integral formula of Snyder et al. (1984). It also uses onsite turbulence data to obtain dispersion coefficients. Thus, we may expect this model to perform better than any of the other models under stable plume impaction conditions.

IMPACT assigns "transparencies" to the horizontal and vertical cell faces in order to accomodate the effects of atmospheric stability. These transparencies were developed on the basis of simulations of idealized problems. It is not known whether or not such assignments are realistic. However, the assignments are based on the PGT classification system which, as mentioned previously, is much too broad a classification scheme to use in predicting plume impaction. The diffusivities used in IMPACT are also determined through the PGT scheme and not through onsite turbulence data.

One class of flows that needs to be dealt with separately is the stable condition where the plume surmounts the hill top. This includes strongly stable flows where the plume is released above the dividing-streamline height as well as moderately stable flows where all the flow surmounts the hill top (dividing-steamline height is zero). This class of flows is not dealt with at all by COMPLEX I or II. COMPLEX/PFM makes the most noble attempt of all the models for treating this type of flow; it allows for a deformation of streamlines (closeness of approach of the plume to the terrain) as a function of stability (Froude number) and terrain shape (crosswind aspect ratio). It is the only Gaussian model that allows for differences in terrain shape, which is known to have a strong influence on the plume trajectory and hence surface concentrations. In principle, COMPLEX/PFM should perform better than the other Gaussian models in this class of flows, although the algorithms for dealing with the adjustments for terrain shape and stability could certainly be improved.

4141 uses a "quarter-height plume" assumption for all stable conditons, irrespective of the plume height or terrain shape. RTDM uses a "half height plume" assumption for all stable conditons where the plume height exceeds the dividing-streamline height. Indeed, when the dividing-streamline height H_D is not zero, RTDM treats the flow as if the ground surface were located at the dividing-streamline height, and the "half-height" is calculated with reference to this pseudo ground surface. This treatment appears to be physically sound (Snyder and Hunt, 1984; Snyder and Lawson, 1984). PLUME5 uses a conservative modification to the half-height assumption. Whereas the RTDM and PLUME5 methods are certainly improvements over the 4141 method, they are clearly not as physically sound as the COMPLEX/PFM method.

EVALUATION OF MODEL PERFORMANCE

The statistical performance measures are quite illuminating and do, in general, illustrate that the soundness of the physics improves model performance.

Many performance measures are presented in the TRC report Wackter and Londergan, 1984) and it is difficult to know which measures are most important. A broad-brush look at all of the measures for the various types of comparisons suggests that the models should be ranked in the following order (best to worst) for the Westvaco comparison:

RTDM
4141
SHORTZ
PLUME5
COMPLEX/PFM
COMPLEX I
COMPLEX II
IMPACT

And for the Cinder Cone Butte comparison:

RTDM
IMPACT
COMPLEX I
SHORTZ
COMPLEX/PFM
PLUME5
4141
COMPLEX II

It should be pointed out that these rankings are not hard and fast. None of the models scored at the same rank for all measures and for all subsets of data. However, generally speaking, rankings based on one measure were surprisingly close to those based on other measures, e.g., if the average difference was small, then the rms error was also small, and the correlation coefficient was generally higher (relatively speaking). Hence, the weighting of the relative importance of the various measures was not highly significant. These performance rankings are essentially independent of the averaging times, as may be expected.

COMPLEX II shows the most consistent and pronounced tendency to overpredict concentrations, for essentially all data sets and all types of comparisons. This presumably results from the overprediction of impingement conditions and overprediction of concentrations under impingement conditions. This is borne out by the tables of Appendix B, where concentrations are grossly overpredicted under stability conditions E and F.

It is not clear why the IMPACT model performed much better (relative to the other models) on the Cinder Cone Butte (CCB) comparison than on the Westvaco comparison. One possible reason is the plume rise algorithm, since plume rise was not a factor in the CCB data but was in the Westvaco data. Much more likely, however, is that the grid covering the Westvaco terrain was much too small to adequately resolve the flow field in this "sea of mountains". Whereas the grid for the CCB terrain allowed a sufficient area (of flat land) surrounding the hill, the Westvaco grid was only 2.6 x 3 km in area and could not allow for the effects of the surrounding high terrain. Also not clear is why 4141 and PLUME5 performed well on the Westvaco comparison, but poorly on the CCB comparison. The most likely reason is that the CCB data are primarily for stable conditions whereas the Westvaco data cover all stability conditions. As pointed out previously, neither of these models allows for terrain impaction, so they might be expected to underpredict concentrations under strong stability. The tables in Appendices B and C, however, do not support this notion.

The fact that COMPLEX/PFM performed better than COMPLEX I and II for the Westvaco comparison suggests that the adjustments for streamline deformations when the plume is above the dividing-streamline height under neutral and stable conditions were worthwhile. The much poorer performance of COMPLEX/PFM compared with COMPLEX I for the CCB comparison, however, suggests that those algorithms for streamline deformations as functions of terrain shape and stability need additional work.

With regard to the predictions of SHORTZ, the user's manual leads the reader to believe that with a ground-based inversion (small mixing depth) and with the plume above the mixing depth, plumes do not contribute significantly to ground-level concentrations at any receptor. The model thus appears to allow impaction under neutral and even unstable conditions (level plume within mixed layer), but not under the most strongly stable conditions. This algorithm is diametrically opposite to current understanding of flows in complex terrain, where plume impaction occurs only under strongly stable conditions. However, the surface mixing layer in this model is defined as the height where the vertical turbulence intensity drops below 0.01. Study of the Modeler's Data Archive from Cinder Cone Butte showed that the vertical turbulence intensity at plume elevation was less than 0.01 in only 1 hour of the total 111 hours. Perusal of one year of Westvaco data also showed that, at typical plume elevations, the vertical turbulence intensity was less than 0.01 only 0.75% of the time (Cimorelli, private communication). Hence, it appears that, in practical terms, the plumes are "always" within the surface mixing layer and hence are allowed to impact on the hills. This is supported by the tables in the appendices, which show that SHORTZ overpredicts most strongly under stability class F. Hence, the end result is that SHORTZ uses similar level-plume trajectory assumptions as do COMPLEX I and II for stable conditions. The fact that SHORTZ performs better than COMPLEX I and II suggests that the use of on-site turbulence data for the computation of dispersion coefficients is helpful.

The comparisons between the predicted and observed concentrations paired in time and location showed significantly smaller discrepancies and significantly higher correlations for the Cinder Cone Butte data set than for the Westvaco data set. I believe this is due to the more refined meteorological data collected at Cinder Cone Butte and to the better control maintained during the experiment. For example, wind directions contained in the Modeler's Data Archive were derived from interpolations of tower wind measurements (to plume elevation) as well as lidar and photographic observations of plume position. Similarly, the single high tower at CCB provided a much better characterization of the relatively homogeneous approach flow there as compared with the one or two measurements per tower at each of three towers scattered over the Westvaco site. Also, plume rise was known (zero) at CCB and unknown (estimated by the models) at Westvaco. These remarks are not intended in any way to denigrate the measurement program at Westvaco, but rather to make the point that the more accurate and comprehensive input data do indeed significantly improve model performance.

I wish finally to comment that the wealth of statistics generated is useful, but so voluminous as to be difficult to digest. Nevertheless, it would be desirable to see even more detailed depictions of the results through various types of graphical displays such as scatter plots in order to isolate the causes of poor model performance. The statistics by themselves only suggest possible causes. Also, the subgroupings are not divided finely enough to isolate particular causes. On the other hand, as pointed out by Wackter and Londergan (1984), it is difficult to select meaningful graphical and tabular displays with a limited report space. But even a modest amount of additional information could further overwhelm a reviewer. I therefore strongly advocate specific case studies as suggested by Irwin and Smith (1984), to evaluate specific strengths and weaknesses in both the data and modeling assumptions. I believe such case studies would be a valuable supplement to the purely statistical approach employed here.

REFERENCES

Briggs, G.A., 1971: Some Recent Analyses of Plume Rise Observations, In: *Proceedings of the Second International Clear Air Congress*, Academic Press, New York.

Briggs, G.A., 1972: Chimney Plumes in Neutral and Stable Surroundings, Atmos. Envir., v. 6, p. 507-510.

Briggs, G.A., 1975: Plume Rise Predictions. In: *Lectures on Air Pollution and Environmental Impact Analyses*, Amer. Meteorol. Soc., Boston, MA.

Hunt, J.C.R. and Snyder, W.H., 1980: Experiments on Stably and Neutrally Stratified Flow over a Model Three-Dimensional Hill, J. Fluid Mech., v. 96, p. 671-704.

Irwin, J. and Smith, M., 1984: Potentially Useful Additions to the Rural Model Performance Evaluation, Bull. Amer. Meteorol. Soc., v. 65, p. 559-568.

Smith, F.B., and Howard, S.M., 1972: Methodology for Treating Diffusivity, Meteorology Research, Inc. (MRI) Publication FR-1020.

Snyder, W.H. and Hunt, J.C.R., 1984: Turbulent Diffusion from a Point Source in Stratified and Neutral Flows around a Three-Dimensional Hill; Part II: Laboratory Measurements of Surface Concentrations, Atmos. Envir., v. 18, p. 1969-2002.

Snyder, W.H. and Lawson, R.E. Jr., 1981: Laboratory Simulation of Stable Plume Dispersion over Cinder Cone Butte: Comparison with Field Data, Appendix, EPA Complex Terrain Model Development First Milestone Report - 1981, Rpt. No. EPA-600/3-82-036, Envir. Prot. Agcy., Res. Tri. Pk., NC, p. 250-304.

Snyder, W.H. and Lawson, R.E., Jr., 1984: Stable Plume Dispersion over an Isolated Hill: Releases above the Dividing-Streamline Height, Appendix, EPA Complex Terrain Model Development Fourth Milestone Report - 1984 (in review), Envir. Prot. Agcy., Res. Tri. Pk, NC.

Snyder, W.H., Thompson, R.S., Eskridge, R.E., Lawson, R.E., Jr., Castro, I.P., Lee, J.T., Hunt, J.C.R. and Ogawa, Y., 1985: The Structure of Strongly Stratified Flow over Hills: Dividing-Streamline Concept, J. Fluid Mech. (to appear).

Strimaitis, D.G., Venkatram, A., Greene, B.R., Hanna, S., Heisler, S., Lavery, T.F., Bass, A., and Egan, B.A., 1983: EPA Complex Terrain Model Development Second Milestone Report - 1982, Rpt. No. EPA-600/3-83-015, Envir. Prot. Agcy., Res. Tri. Pk., NC, 375p.

Wackter, D.J. and Londergan, R.J., 1984: Evaluation of Complex Terrain Air Quality Models, Rpt. to Envir. Prot. Agcy. under Contract No. 68-02-3514, Res. Tri. Pk., NC, 233p.